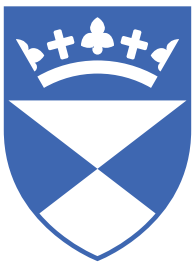
## DOCTOR OF PHILOSOPHY

# Computer-Assisted Analysis of Arterial Narrowing in Whole-Body Magnetic Resonance Angiography

McNeil, Andrew

*Award date:*
2020

# Computer-Assisted Analysis of Arterial Narrowing in Whole-Body Magnetic Resonance Angiography



University of Dundee

**Andrew McNeil**

University of Dundee
United Kingdom

This dissertation is submitted in partial fulfilment for the degree of *Doctor of Philosophy*

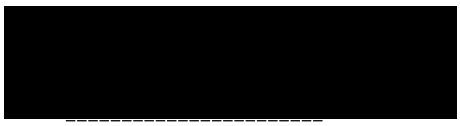
School of Science and
Engineering (Computing)

June 2020

# Declaration of Authorship

## Candidate's Declaration

I, Andrew McNeil, hereby declare that I am the author of this thesis; that all references cited have been consulted by me; that the work of which this thesis is a record has been done by me, and that it has not been previously accepted for a higher degree.



Signed

05/06/20

Date

## Supervisor's Declaration

I, Emanuele Trucco, hereby declare that I am the supervisor of the candidate, and that the conditions of the relevant Ordinance and Regulations have been fulfilled.

Signed

05/06/20

Date

# Acknowledgements

I would first like to thank my supervisor Prof Emanuele Trucco for his continual support, guidance, and encouragement over the course of this project. I have learned much over these years, not just about being an effective and critical researcher, but also a constructive and supportive supervisor, which I have tried to emulate in my own supervision of undergraduate students, and will carry forward into my future career.

My thanks also go to my other supervisors; Prof. Graeme Houston at Ninewells Hospital provided indispensable clinical support for this project, and his insight and expertise in guiding our work was crucial to its success. And my industry supervisor, Dr Ian Poole at Toshiba Medical Visualisation Systems (now part of Canon Medical) provided much needed guidance and a critical eye on the direction of our research, and his support during the year I spent on placement within the company gave me an invaluable look at the "other side of the fence" of industry-led research.

I am also grateful to Medical Research Scotland whose financial support made this project possible, and whose informative training sessions every year helped expose me to a much wider variety of biomedical and pharmaceutical research.

It was a pleasure to be a part of the CVIP group at Dundee and the collaborative, supportive atmosphere that it fosters. I would like to thank Prof. Stephen McKenna, Dr Jianguo Zhang, and all of the postdocs and PhD students who have been a part of it during my time there. The workshops, reading groups, and weekly meetings all helped hone both my abilities as a researcher, the direction of my work, and my confidence in it.

And finally, a very special thank you goes to my family and my wonderful wife, whom I had only just met a few weeks before the start of this endeavour. Their unwavering support and encouragement kept me motivated, helped me through the hard times, and pushed me forward to achieve everything I could.

# List of Publications and Achievements

## 1a. Publications from Project

1. **Andrew McNeil**, Giulio Degano, Ian Poole, Graeme Houston, Emanuele Trucco. "Comparison of Automatic Vessel Segmentation Techniques for Whole Body Magnetic Resonance Angiography with Limited Ground Truth Data", Medical Image Understanding and Analysis: 21st Annual Conference, MIUA 2017, 144–155. (2017)

2. Lynne McCormick, Jonathan Weir-McCall, Stephen Gandy, Richard White, **Andrew McNeil**, Emanuele Trucco, J Graeme Houston. "Mapping of arterial location for the design of automated identification and analysis algorithms in whole body MRA", 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, 2015, pp. 7925-7928.

## 1b. Additional Publications

1. Paul F. Brennan, **Andrew J. McNeil**, Min Jing, Agnes Awuah, Dewar D. Finlay, Kevin Blighe, James A.D. McLaughlin, Ruixuan Wang, Jonathan Moore, M. Andrew Nesbit, Emanuele Trucco, Mark S. Spence, Tara C.B. Moore, "Quantitative assessment of the conjunctival microcirculation using a smartphone and slit-lamp biomicroscope", Microvascular Research, 103907 (2019).

# 2. Unrefereed Papers and Presentations

### College of Medicine Symposium, Dundee University

*Poster Presentation*: "GroBa: A lumen calibre measurement system for MRA Data". February 2014.

### SINAPSE Annual Scientific Meeting, Edinburgh

*Poster Presentation*: "GroBa": A system for lumen calibre measurement in WBMRA". June 2014.

### Dundee Medical Image Analysis Workshop

SICSA-sponsored medical image analysis workshop, held in the Computing department at Dundee University.
*Poster Presentation*: "GroBa: "Growing Balloons" for lumen calibre measurement in WBMRA". March 2015.

### Medical Research Scotland Commercial Training Event

These two-day events, ran in March every year, brought together all MRS students for talks from industry professionals, as well as presentations and posters from other students.

- 2014 - *Oral Presentation*: "Computer-Assisted Analysis of Arterial Narrowing in Whole-Body Magnetic Resonance Angiograms (WBMRA)".
- 2015 - *Poster Presentation*: "Computer-Assisted Analysis Of Arterial Narrowing In Whole-Body MRA".
- 2016 - *Oral Presentation*: "Joining the Organisation Committee for a Medical Image Analysis Workshop".
- 2017 - *Oral Presentation*: "Towards an Automated System for Detecting Arterial Stenoses in Whole Body MRA".

# 3. Training Events and Achievements

### Medical Imaging Summer School 2014, Favignana, Sicily

Attended MISS'14 summer school (28th July – 1st August 2014), where the courses (totalling 24 hours) covered both theoretical and practical aspects of Medical Imaging and Computer Vision, as well as examples of their commercialisation.
*Poster Presentation*: "GroBa": A system for lumen calibre measurement in WBMRA".

### PhD Symposium, School of Computing, Dundee University

Two day symposium, held yearly, in which all Computing PhDs submitted a 2-page abstract of their research, reviewed by two peers, and delivered a 20 minute presentation on their work to other students and professors.
Awarded **Best Paper Award for Outstanding Doctoral Student**, 2017.

### CVIP Technical Workshops

Attended and presented regularly at the twice-yearly CVIP technical workshops throughout the PhD (2013-2019).

### Dundee Medical Image Analysis Workshop

I helped co-organise the SICSA-sponsored workshop, including collating and preparing the book of abstracts for the event.

### Careers Fair, Royal High School, Edinburgh.

I presented my research to S4 and S5 pupils at careers fair, held at Royal High School, Edinburgh, September 2016.
*Oral Presentation*: "A Career in Science Research: Planning for University and Beyond...".

# Abstract

Cardiovascular disease (CVD) is one of the leading causes of death and disability in the UK. There is therefore much interest in the early staging of CVD to help improve patient outcomes through long-term treatments. Whole-body magnetic resonance angiography (WBMRA) offers a quantitative assessment of whole body arterial atheromatous disease burden, making it a useful diagnostic tool for CVD. However the large datasets produced are very labour-intensive to examine, and there is an unmet need for automated software tools to help clinicians with their diagnoses. Towards this goal, this thesis proposes an automatic framework for the processing and analysis of WBMRA data for grading stenoses. This includes registration of pre- and post-contrast volumes, automatic vessel segmentation and tracking, and the measurement and grading of tracked vessels.

We present the first quantitative comparison of vessel segmentation techniques for WBMRA data, comparing five different algorithms using 3 ground truth vessel maps annotated manually following a clear protocol. We find that a U-Net convolutional neural network algorithm outperforms previous, well established algorithms despite the limited amount of training data.

To enable the development and validation of stenosis grading algorithms, we also gathered ground truth stenosis annotations for 18 patients with three trained clinicians. A thorough analysis of the inter- and intra-rater variability revealed a higher disagreement between annotators than expected for manually detecting and grading stenoses, which is largely unexamined in the literature. The development of clear protocols for the collection of this ground truth data, in close collaboration with clinical partners, enables us to present good-practice guidelines for ground truth collection from WBMRA data for algorithm development.

Finally, we present three stenosis detection algorithms tested against synthetic vessels and well-characterised ground truth stenosis annotations.

# Contents

# List of Figures

# List of Tables

# Acronyms

**CEMRA** Contrast-Enhanced Magnetic Resonance Angiography.

**CNN/Convnet** Convolutional neural network.

**CT** Computed Tomography.

**CTA** CT Angiography.

**CVD** Cardiovascular disease.

**DSC** Dice similarity coefficient.

**MACE** Major adverse cardiovascular event.

**MRA** Magnetic Resonance Angiography.

**MRI** Magnetic Resonance Imaging.

**WBMRA** Whole-Body Magnetic Resonance Imaging.

# Chapter 1

# Introduction

This thesis focusses on the automated analysis of arteries in whole-body magnetic resonance angiography (WBMRA) data, with the particular aim of developing a system to assist clinicians diagnose stenoses in the arterial lumen. This work was funded by Medical Research Scotland and Toshiba Medical Visualization Systems Europe (now Canon Medical Research Europe Ltd.), with the work carried out in collaboration with Canon Medical and the Cardiovascular and Diabetes Medicine group at Ninewells Hospital, Dundee, UK. Canon provided technical assistance and training, with 12 months spent on placement within the company over the course of the project, and the clinical team providing patient data and ground truth annotation expertise.

In this chapter we will cover the background and motivation for arterial analysis in WBMRA, from both a clinical and technical perspective, before summarising the key contributions of this thesis and concluding with an overview of the thesis organisation.

## 1.1 Background and Motivation

Each year cardiovascular disease (CVD) causes over 3.9 million deaths in Europe; accounting for 45% of all deaths [10]. CVD is also the main cause of death in men in all but 12 countries of Europe and is the main cause of death in women in all but two countries. This places a large burden on society, with an estimated cost to the EU economy of almost €210 billion a year due to health care costs, productivity losses and the informal care of people with CVD [10]. There is therefore much interest in the early staging of CVD (identifying its severity and distribution, similar to the grading of cancer) to improve patient outcomes through long-term lifestyle changes and treatments.

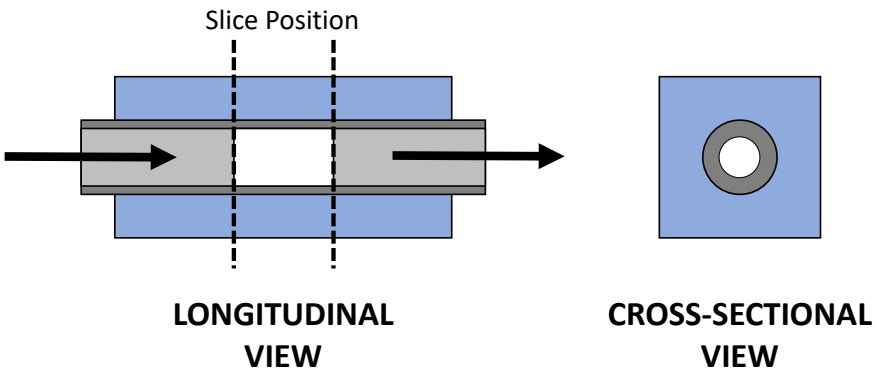



**Figure 1.1** – Schematic of the contrast bolus tracking technique, where the image is acquired as the contrast bolus (shown in white) passes through the plane of interest (black dashed lines), generating high contrast in the vessel lumen (shown in light grey). (Adapted from [1])

Major adverse cardiovascular events (MACE) is a composite endpoint with varying definitions which is commonly used in cardiovascular research, and generally encompasses nonfatal stroke, nonfatal myocardial infarction, and cardiovascular death [11]. Whole-body magnetic resonance angiography (WBMRA), sometimes called whole body cardiovascular magnetic resonance angiography (WB-CMRA), offers a quantitative assessment of whole body arterial atheromatous disease burden in a single non-invasive imaging study. It has been shown to be an effective tool for the stratification of cardiovascular disease [3], with Total Atherosclerotic Score (TAS — a measurement of the overall atherosclerotic burden of the arterial tree derived from the severity scoring of 26 main arteries) serving as a risk predictor of MACE independent of major cardiovascular risk factors [12].

This validated quantitative score of whole body atheroma burden and individual arterial disease sites allows identification of patients by site, overall severity and burden of disease, and by the likelihood of rapidly progressive disease which could guide clinical management planning. The use of the whole body approach in case selection can aid in determining which site may be most likely to show a beneficial effect of intervention such as revascularisation or medical therapy, and aids in detecting patients with more rapidly progressive disease.

### 1.1.1 Imaging Modality

In contrast-enhanced Magnetic Resonance Angiography (CEMRA), the contrast in the arterial lumen is enhanced by injecting a contrast agent into a vein in the arm and acquiring images using an MRI scanner as the agent passes through the arteries of interest [1]. This technique gener-

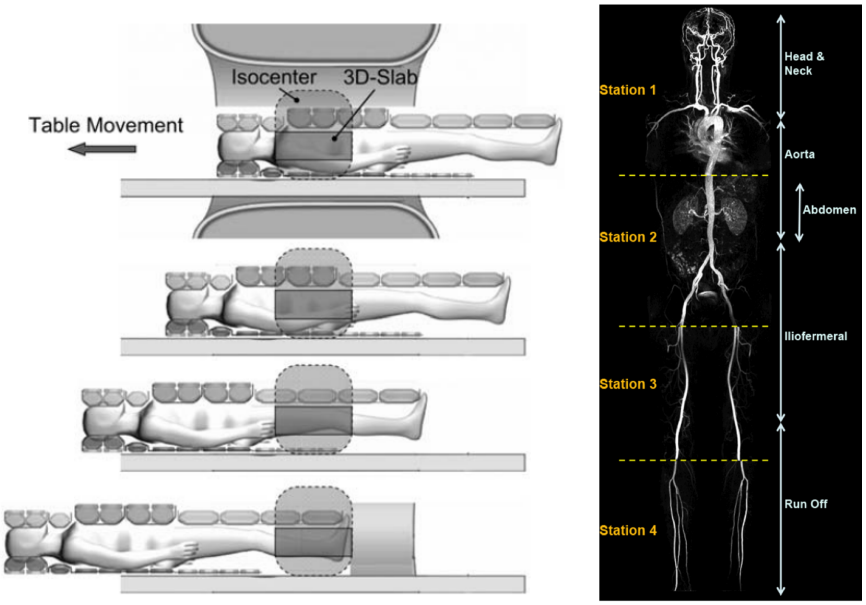



**Figure 1.2** – (L) Schematic of WBMRA acquisition, showing image volumes acquired at different stations with the patient table moved between each scan [2]. (R) Example stitched image showing how the field-of-view of each station. (Source: [3])

ates high contrast in the lumen (the channel where the blood is flowing), providing a non-invasive, comprehensive imaging method for assessing cardiovascular disease (CVD) throughout the entire body [13]. The critical factor in generating high contrast is to time the scan acquisition to coincide with the peak contrast agent concentration in the area of interest (Figure 1.1). The duration of the bolus peak may be less than the duration of the sequence itself due to dispersion over time, with a previous study showing a peak time of just 7 – 13 seconds in the aorta [14]. If the scan is mistimed this can lead to either poor arterial enhancement if the scan is too early (often an issue in the extremities of the legs where the optimal timing becomes harder to predict), or unwanted venous enhancement if it is triggered too late, leading to spurious enhanced vessels in the dataset which are hard to distinguish from arteries.

Whole body scans are acquired in a number of stages, or body segments (head, abdomen, legs, feet), with the patient table moved between each scan (Figure 1.2). The individual volumes can then be stitched together to give an overall view of the entire body. Typically two scans of patient are acquired, pre- and post-contrast injection, allowing the two to be subtracted to suppress static tissues and make the vessels more easily observable.

The main challenge of these studies is the large imaging dataset which is clinically time-consuming to examine manually in detail. The devel-

opment of robust, automated quantitative analysis tools to help stage CVD from WBMRA examinations and aid reporting of WBMRA — something which is not currently satisfied by any existing software solution — would significantly assist in the clinical workflow and encourage more widespread adoption of this technique for CVD assessment.

### 1.1.2 Example Clinical Diagnostic Workflow

The general outline of the typical diagnosis protocol at Ninewells Hospital for CVD using MRA, from when the patient is admitted to the hospital until the decision for treatment and/or intervention is made, is as follows:

1. The patient is referred to a radiology department with recorded symptoms and request for a diagnostic scan.

2. A consultant radiologist determines the optimal scan routine to be carried out based on symptoms and suspected diagnosis (such as contrast-enhanced MRA for CVD quantification).

3. The patient is booked for a scan, which is performed using the recommended scan routine.

4. The resulting data volumes are inspected by a qualified radiologist for signs of CVD. This is typically done manually by examining slice-by-slice in the axial, sagittal and coronal directions, with narrowings in the lumen quantified by eye.

5. Results of the MRA assessment is discussed at a Multidisciplinary Team Meeting (MDT), where treatment and possible intervention is decided.

### 1.1.3 Existing Software Solutions

A number of commercial systems exist which aim to assist the diagnostic process by providing semi-automated tools for vascular segmentation and stenosis assessment:

**Withinsight Automation Framework** by Claron Technology Inc. [15], a multi-modality tool offering atlas-driven segmentation and analysis and automatic whole-body vessel extraction.

**COR Analyzer** by Rcadia [16], a system for analysing Coronary CT Angiography (cCTA) studies to identify significant stenotic lesions

in the coronary arteries, and has been given FDA clearance for ruling out Coronary Artery Disease in coronary branch vessels.

**VesselMASS RE** by Medis Specials [17], a research tool for quantifying vessel wall dimensions and plaque composition from MR images.

All of these typically require a high level of manual intervention and user training, and the quantification tools may be restricted to certain areas of the body. So whilst there have been a number of publications relating to these tools, their use is still largely limited to providing a decision aid as a second reader [18, 19], and they have seen very limited adoption into clinical and diagnostic settings.

The focus of our work presented here was therefore to develop a fully automated framework, requiring no manual intervention at any stage of the pipeline, which both segments the full arterial tree from WBMRA examinations and detects areas of stenosis within it. This framework was developed in close collaboration with clinicians, guided by the demands of the diagnostic task.

## 1.2 Contributions

In this thesis we put forward a fully automated framework for the processing and analysis of WBMRA data for grading stenoses. We contribute to the existing literature of automated analysis of WBMRA through the quantitative comparison of vessel segmentation techniques, validated against manual ground truth, and through quantifying the inter- and intra-observer variability of manual stenosis grading by trained observers, providing a clear protocol and recommended guidelines for collecting annotations to be used as ground truth data for algorithm development. The key contributions of this thesis can be summarised as follows,

1. First quantitative comparison of vessel segmentation techniques for WBMRA data, including the first application of a 3D convolutional neural network to the task of automatic vessel segmentation in WBMRA.

2. Quantification of the inter-observer variability for the task of manually detecting and grading stenoses from WBMRA examinations by trained clinicians.

3. Development of clear protocols for the collection of ground truth data for vessel segmentation and stenosis severity in WBMRA vol-

umes, putting forward good-practice guidelines for ground truth collection from WBMRA data for algorithm development.

4. First fully automated framework for extracting and analysing arteries in WBMRA examinations to be tested against well-characterised ground truth stenosis annotations in real patient data.

## 1.3 Thesis Organisation

Figure 1.3 shows a schematic overview of the thesis as it relates to the developed processing pipeline.

**Chapter 2: Related Work.** We review work related to the clinical use of WBMRA for CVD diagnosis, with a concise review of technical approaches to WBMRA analysis.

**Chapter 3: Clinical and Technical Motivation.** We first describe a typical workflow which is followed when examining WBMRA data clinically. We then describe a "paper-prototyping" exercise undertaken in close collaboration with clinicians, which highlights the clinical demands required of a diagnostic tool and the technical developments which must be pursued in order to fulfil that unmet demand.

**Chapter 4: Materials and Preprocessing Methods.** We describe the image data used in our work, and any preprocessing methods which were common to all experiments using the data.

**Chapter 5: Data Pre-Processing and Baseline Method Evaluation.** Here we examine the baseline pipeline on which this project builds. This includes a qualitative comparison of two off-the-shelf volume registration methods, and evaluation of a vessel calibre measurement algorithm as described in [9].

**Chapter 6: Automatic 3D Vessel Segmentation.** We present the first quantitative comparison of vessel segmentation methods for WBMRA data, comparing three previous approaches with our own 3D Convnet and 2D U-Net algorithms.

**Chapter 7: Ground Truth Stenosis Assessments and Variability Analysis.** The data collection exercise undertaken to collect ground truth stenosis annotations is analysed in detail here. We

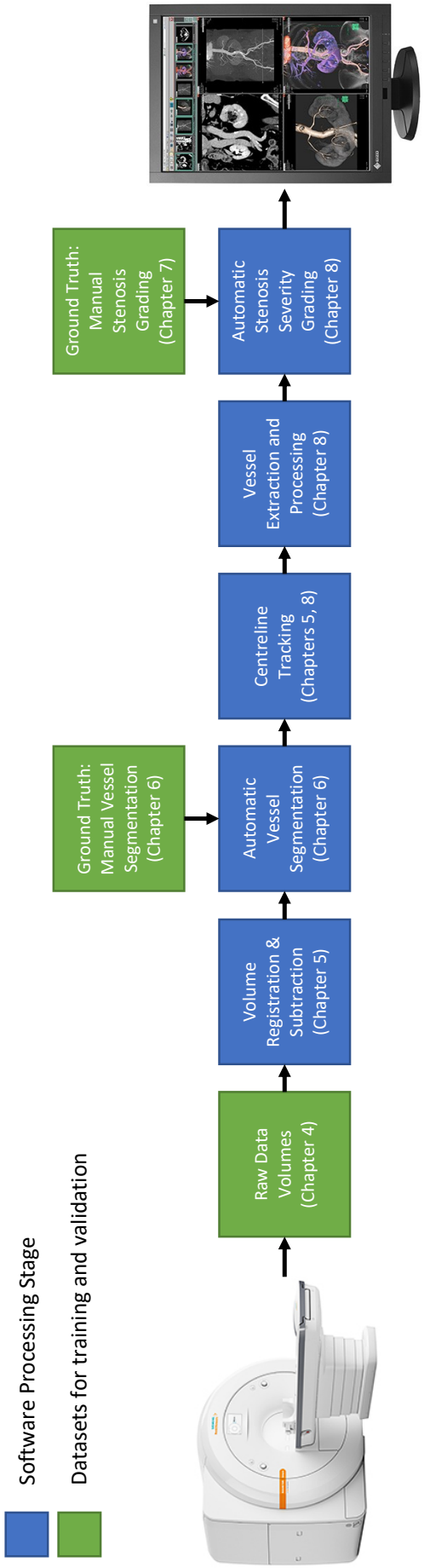



**Figure 1.3** – Schematic overview of thesis layout.

describe the protocol developed, and examine the statistical differences between annotators and the impact this may have on any algorithms developed using such data.

**Chapter 8: Automated Stenosis Grading.** This chapter presents the results of applying our developed pipeline, as described throughout preceding chapters, to the task of detecting and quantifying stenoses. We then present three stenosis detection algorithms, applied firstly to a large number of synthetic vessels with artificial stenoses of known severity, then to the real data described in Chapter 7.

**Chapter 9: Conclusions, Discussion, and Future Work.** The final chapter summarises the main conclusions of this work, and suggests future directions to be explored in order to build upon the results we have presented.

# Chapter 2

# Related Work

## 2.1 Introduction

There exists a substantial volume of literature dedicated to the general task of extracting and measuring vessel-like structures in volumetric datasets [20, 21]. This stems from the fact that many techniques, despite having modality-dependant aspects, can generally be applied to many modalities. As such, methods developed for CTA can typically be applied to MRA, with some techniques being 3D extensions of previously derived 2D techniques from other areas. In this chapter we will therefore focus on the main areas relating to the technical work carried out in developing our processing pipeline and algorithms.

We will begin with a brief discussion of the imaging modality of WBMRA, how it is used for CVD quantification, and the importance of ground truth annotations for algorithm development. We will then examine three key areas central to developing our automated framework; vessel segmentation, centreline extraction and vessel calibre measurement. Finally we will examine the task of stenosis detection using these techniques.

This chapter aims to give an overview of the current literature relating to vascular analysis in WBMRA. Detailed descriptions of the techniques used in our technical developments will be reserved for their appropriate chapters.

## 2.2 WBMRA for Cardiovascular Risk Assessment

### 2.2.1 Imaging Modality: WBMRA

MRA is a non-invasive imaging modality that has been widely used for stenosis detection and screening [22–24]. Early diagnosis of secondary complications from, and better treatment planning of, Diabetes Mellitus (DM) is also an area of active interest [25]. Indeed, it has recently been shown that Whole-Body MRA can provide long-term prognostic information regarding future cardiovascular events in patients with DM [26].

One of the major interests in future applications of WBMRA is the potential for screening asymptomatic populations [25], where the use of no ionizing radiation makes it safe and non-invasive, and scan times of around an hour gives good patient tolerance. There are however cost-effectiveness concerns for screening due to low expected prevalence of significant cardiovascular pathology in asymptomatic populations, and the fact that the large datasets are labour-intensive (and therefore expensive) to examine [25].

The image contrast in an MRA image is highly dependant on the pulse sequence used, and the greylevel values in the image are therefore not directly relatable to a physical property of the object, as in X-Ray CT for example [27]. There are also a number of artefacts common to all MRI techniques, such as bias field, close proximity to receiver coils (particularly in the legs) causing signal loss, patient movement, etc. [27], in addition to artefacts specific to WBMRA. For example the multi-station technique requires acquisition from several discrete, easy-to-overlap FOVs, so gradient distortions on the margins of the individual FOVs can potentially lead to discontinuity artefacts [2].

### 2.2.2 Cardiovascular Risk Assessment

CVD is a broad term that encompasses four conditions [28],

1. Coronary heart disease: occurs when the blood flow to the heart is restricted due to a build-up of fatty substances in the coronary arteries.

2. Stroke: occurs when the blood supply to a region of the brain is impeded and is a medical emergency.

3. Peripheral arterial disease (or peripheral vascular disease): occurs when the blood supply to the limbs becomes restricted.

4. Aortic disease: occurs when the aorta swells causing an aneurysm. The most common is an abdominal aortic aneurysm (AAA) often diagnosed by an ultrasound scan.

CVD risk assessments are typically carried out for the most at-risk age group of the over 40 – 79. Initially, the risk of CVD is estimated using the person's age, sex, ethnicity, family history of CVD, blood pressure, level of cholesterol, blood sugar levels, physical activity levels and whether or not they are a smoker [28]. Guidelines vary, but the American College of Cardiology recommends calculating the 10-year risk for heart disease every 4 – 6 years in patients 20 – 79 years old who are free from heart disease and 30-year risk of developing heart disease for patients 20 – 59 years of age who are free of heart disease and are not at high short-term risk for heart disease [29]. This is so that effective long-term lifestyle and dietary treatments can be recommended early to reduce risk factors where needed. For patients older than 79 years of age without known CVD, it is recommended to discuss ongoing risks and benefits of primary preventive therapies where needed, as it is unknown at what age periodic risk assessment should no longer be performed due to many of the validated risk models only including patients up to 79 years of age [29].

The presence of peripheral arterial disease (PAD) is often a pre-cursor to other cardiovascular conditions, such as coronary heart, disease, stroke and aortic disease. Peripheral arterial disease is typically diagnosed via a physical examination by a general practitioner. The ankle brachial pressure index (ABPI) is often used to diagnose PAD and is calculated by comparing the blood pressure in the upper arm and the ankles [30]. If the patient has PAD the blood pressure in the ankle will be lower due to the obstructions to the blood flow. Further testing by means of a radiology scan (MRA or CTA) may be required if, for example, the ABPI score is normal but the patient expresses PAD symptoms or the arterial disease is so advanced that it requires surgical intervention.

Atherosclerosis is a systemic disease, thus absence of plaque in one territory cannot be equated to absence of plaque in another [31]. WBMRA offers a systemic quantification of the global atherosclerosis burden within the body, and the atheroma score from WBMRA has been shown to correlate well with traditional cardiac risk factors, the prevalence of ob-

structive coronary artery disease and future major adverse cardiovascular events [31–33].

### 2.2.3 Ground Truth Annotations

One key issue which is central to all areas of medical image analysis is that of obtaining ground truth annotations. Knowing the underlying truth, whether it be the true medial axis of a vessel or the location of a stenotic lesion, is essential for the development and validation of image processing algorithms. However, there are often a number of issues which complicate this process. The acquisition of such annotations is typically very time consuming, and for many tasks it may be prohibitively expensive or even not entirely possible to acquire the actual gold standard labels for training [34]. There is also often significant inter- and intra-observer variability, even between experienced clinicians [35, 36]. This can place constraints on the accuracy of a potential algorithm due to either lack of training data, or because there is no accurate consensus on which to validate the algorithm's performance.

A number of attempts have been made to tackle these issues. In the case where there is a lack of a gold standard ground truth and instead there are multiple noisy labels provided by multiple annotators, between whom there may be significant disagreement, a probabilistic approach has been put forward for supervised learning [34]. In [34] a Bayesian framework is used to create an algorithm which iteratively establishes a particular gold standard, measures the performance of the annotators given that gold standard, and then refines the gold standard based on certain performance measures, the results of which proved superior to the commonly used majority voting baseline [34]. Another approach has involved using deep learning to determine the spatial image features needed for training a supervised classifier from a large amount of unlabelled data [35]. An approach developed at tackling the issue of the time and cost involved in gathering ground truth is to use a large number of annotations gathered from non-expert volunteers [37, 38], where it has been found that for some applications label noise can be compensated by a sufficiently large corpus of training data [39]. However, these approaches must achieve high reliability to meet the definition of "ground truth" [40, 41]. There has also been recent work focussed on the automatic generation of annotations for non-medical classifiers with large numbers of classes [42], and there is growing interest within the medical image analysis community for techniques which allow a drastic

reduction to the typical number of annotations required at a parity of performance [43].

## 2.3 Vascular Analysis

In this section we will discuss the methods used in the two key steps of most vascular analysis frameworks; vessel segmentation followed by vessel tracking via centrelines.

### 2.3.1 Vessel Segmentation

When using a subtracted CEMRA dataset, the first stage is often to enhance the vessel-like structures within the data. This has the effect of not only enhancing the structures of interest, but also suppressing the signal from other anatomical features which aren't of interest (e.g. artefacts left after subtraction due to motion in the abdominal cavity).

One common technique employed is to look at the local curvature by analysing the Hessian (local second derivative) [44, 45]. The main concept underpinning eigenvalue analysis of the Hessian is to extract the principal directions in which the local second order structure of the image can be decomposed. From this, the direction of smallest curvature (along the vessel) can be determined. For the application of enhancing vessels in a 3D angiographic dataset, a "vesselness" measure is analysed at different scales, defined as a combination of three measures which differentiate between blob-like structures, plate-like structures, and background noise [44]. The response of this filter will be maximum at a scale which approximately matches the size of the vessel being analysed. Therefore, the final enhanced map is generated by integrating the vesselness measure provided by the filter response at different scales, which can then be thresholded to produce a binary vessel tree.

Despite being a well established tool in both 2D and 3D vascular analysis, the multi-scale Frangi vesselness filter described above does have some drawbacks. The main issue is that because it is only concerned with elongated vessel-like structures it tends to give a weak response around vessel junctions and bifurcations, often resulting in disconnected branches [46]. Some recent work has been targeted at improving this limitation in the case of 2D imaging using multi-scale invertible orientation scores, obtained by correlating the input image with a specially designed anisotropic wavelet, and applying the vesselness filters in this domain [47]. Other filters have also been developed for the same purpose

such as Optimally Oriented Flux (OOF), first reported in [48], which evaluates a scalar measure of the flux flowing through a spherical surface. Before computing this value, directional information is extracted by projecting the gradient along "optimal" axes, and the flux then evaluated. For each voxel a sphere with variable radius is built, centred on the voxel, which produces an "OOF response" when touching an object edge, and giving a positive response when inside a curvilinear structure. As with Frangi, the goal is to obtain the principal eigenvalues for each voxel; an enhanced map is obtained by the maximum response of the OOF filter evaluated at appropriate range of scales. This approach was found to give a more robust enhancement of vessel like structures in medical images, particularly around vessel junctions [49].

Another approach has been explored in the area of vessel enhancement, where scale space theory has be utilised for smoothing and enhancing medical images by employing the diffusion equation; the original image is the initial condition of the function, and the diffusion function is specified either as scalar or tensor based. One such approach developed for MRA is based on the combination of a non-linear diffusion filter with and Expectation Maximization (EM) technique. The non-linear diffusion filter smooths the homogeneous regions while preserving edges, whilst the EM technique finds the optimal statistical parameters based on the probability distribution of the classes to discriminate the tissues in the image. In this way the optimal contrast for enhancing vessels is found, smoothing and dimming the embedded tissues around the vessels while brightening the vessels themselves [50]. This method performed considerably better than previously used nonlinear diffusion filter approaches, though a direct comparison with the simpler Frangi filter was not made [50].

Another commonly used technique for vessel segmentation is the application of level sets [51–54]. The key idea of level sets is to transform the potentially complex evolution of one or multiple curves (e.g. a vessel cross-section) over time, into the zero level of a higher dimensional surface which evolves based on a potential field (or speed function) [55]. Figure 2.1 shows a schematic of how the surface and zero level set evolves over time. One of the key advantages of this method is that it can very easily describe shapes that change topology, e.g. when a shape splits in two, such as a vessel bifurcation.

Most recent advances in medical image segmentation however have been focussed on the use of convolutional neural networks. Deep CNN

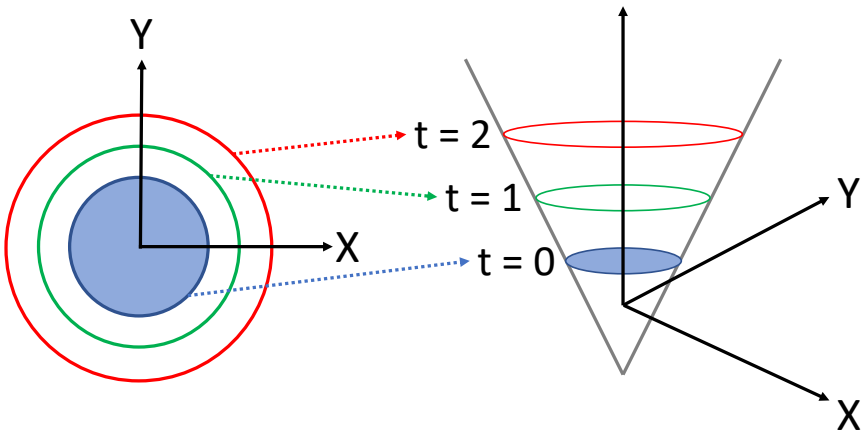



**Figure 2.1** – Transformation of front motion into the level set framework, where it is now an initial value problem. Here we can see how the zero level set evolves over time.

approaches have been driving advances in many computer vision tasks, such as image classification [56,57] and image segmentation [58,59]. Many network models have been developed for these tasks, and it is a very active area of research [60]. For medical image segmentation, the most successful network structures are based on the "U-Net", first proposed by Ronneberger *et al.* [58]. Such deep networks typically require large amounts of data to be train however, and many methods are developed primarily for 2D images due to GPU memory constraints, and lack of public 3D ground truth datasets with which to train. Prior to the work presented in this thesis, no deep learning methods had been applied to the task of vascular segmentation in WBMRA.

### 2.3.2 Vessel Tracking

The extraction of the vessel centrelines is a pre-processing stage which in the case of three-dimensional angiography aims to reduce the 3D vessel to a simpler one-dimensional representation, while still preserving its topology. Centrelines are used in a number of applications, such as registration, anatomical segmentation, stenosis quantification, and visualization and surgery planning (e.g. vessel "fly-throughs" [61]).

The simplest methods rely on a binary segmentation of the vessels, such from simple thresholding, and define the centrelines as the skeleton of that image. One approach is to extract the skeleton through topological thinning, iteratively removing boundary voxels whose removal does not alter the object topology (i.e. retains the same connectivity as the original object and does not introduce any breaks or cavities). One such algorithm [62] involves first computing the distance transform of the ob-

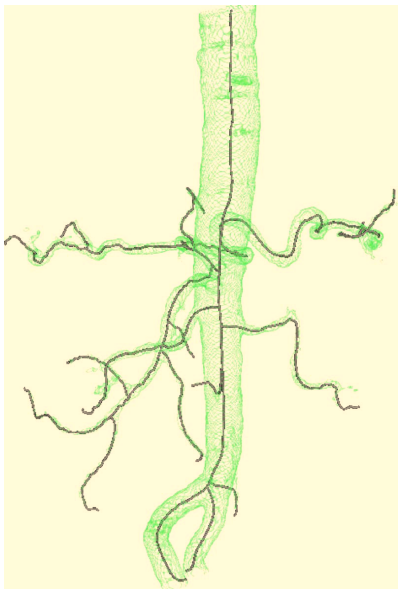


**Figure 2.2** – Binary segmentation of descending aorta, renal, and mesenteric arteries, with corresponding subvoxel precise skeleton. (Source: [4])

ject [63], followed by the divergence of the distance transform gradient using the average outward flux in the neighbourhood of the voxel. Finally, the binary image is thinned by removing the simple points ordered by decreasing divergence. This skeletonization process is not invariant to the order of the removal, hence the need to remove in the order of decreasing divergence, and is typically followed by a post-processing smoothing step [62].

One of the common issues encountered with some approaches is their difficulty in dealing with abrupt variations in local artery direction and sudden changes of lumen calibre, such as around stenoses. A number of algorithms therefore make use of additional information, such as the eigenvalues of the Hessian matrix, in order to identify the true vessel points more reliably [64, 65].

One approach which aims to get around the problem of many algorithms which require a significant amount of user interaction, and also calculate a discrete and often coarse representation of the vessel by using a voxel grid, is based on a number of fast marching method propagations to extract the skeleton at subvoxel precision [4]. By using a level set method, a subvoxel precise distance field is used as an input, thereby removing the skeleton from the constraint of being on the voxel grid and hence generating subvoxel precise centrelines (Figure 2.2).

Finally, some methods also determine the centreline implicitly as part of the segmentation process, such as in the case of snakes (a form of ac-

**Figure 2.3** – Finding the centreline of the vessel by region growing with a volumetric limitation. (Source: [5])

tive contour) [5]. Here the lumen is segmented through iterative region growing, constraining both the number of voxels added in each iteration and limiting the data range to the expected level of the lumen. The centre of the volume at each step is taken as a control point for a CR-spline, eventually approximating the vessel centreline by integrating all the control points (Figure 2.3). Finally, a second refinement step is used where additional 2D region growing is applied in the cross-sections perpendicular to the extracted centreline, adjusting the control point to the centre of the segmented area [5].

## 2.4 Vessel Calibre Measurement and Stenosis Estimation

In this section we will discuss the algorithms used for measuring the calibre of tracked vessels in 3D data, and the methods developed for detecting potential stenoses from these measurements.

### 2.4.1 Parametric Model Fitting

Parametric models try to fit a tube-like model into the lumen, exploiting the fact that vessels are elongated and roughly tubular [66, 67]. These have been shown to work well on regular and uniform vessels, but not in the presence of stenoses where the cylindrical model isn't suited to picking up subtle calibre variations, and where natural anatomical variations deviate significantly from a cylinder. These methods also typically require human intervention to label the start and end points of the vessel, and to correct the fitting where it falls into a local minima.

One such example is the method for automated segmentation of the carotid artery lumen from MRA using a deformable tubular 3D Non-

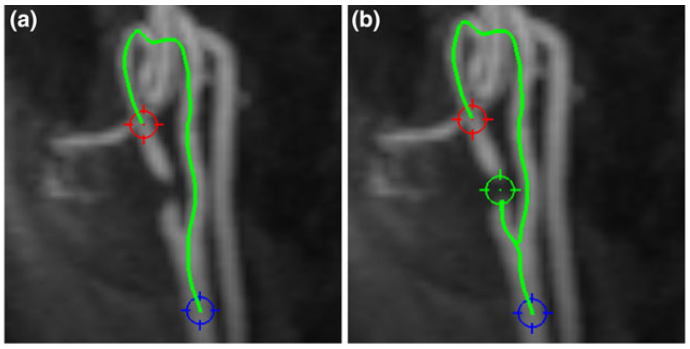


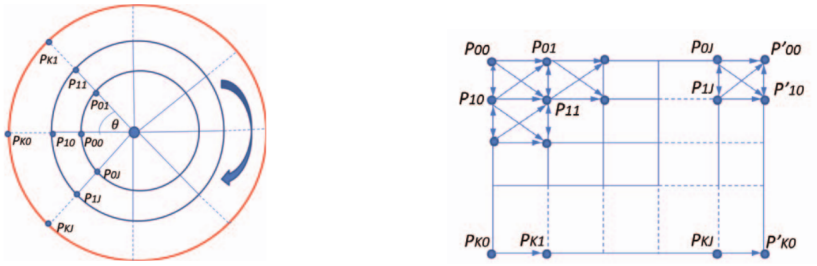

**Figure 2.4** – (a) Failed pathline detection in the presence of a severe stenosis. The path takes a route around the stenosis, crossing into another close vessel, rather than tracking across the stenosis. (b) even after an intermediate point (green marker) was added to help the pathline detection, the algorithm still fails to track across the stenosis. (Source: [6])

Uniform Rational B-Splines (NURBS) model, proposed in [6]. Here, a fast marching level set method was used to determine the centreline along which the NURBS model was initiated, with a gradient-based energy minimization iteration scheme being used to deform the tubular model within the lumen. However, in areas of complex vascular shapes manual intervention was required to correct the pathline, and in the case of a severe stenosis even the manual corrections failed to improve the result (Figure 2.4).

### 2.4.2 Cross-Section Based Measurements

Cross-section based methods rely on finding the longitudinal axis (the centreline) of the lumen and obtaining the perpendicular cross-section, giving the lumen area and calibre at each point along the axis. However, small errors when estimating the centreline often lead to inaccurate calibre measurements [68].

A graph-based method has been proposed to find the optimal cross section boundary for vessel segmentation and stenosis quantification [7]. It was motivated by the problem that centreline-tracing approaches for vessel segmentation incorporate the radius with the 3D voxel coordinates, giving an approximate surface of the vessel when the tracking ends. However, this assumes a circular profile to the vessel which doesn't hold true in the presence of stenoses. The approach proposed by Zhu *et al.* finds the optimal cross section boundary by transforming the cross section voxels to nodes on a graph (Figure 2.5), thereby converting the optimal boundary problem to finding the optimal path in the graph. However, the orientation of the cross section is still highly dependant on the centreline,




**(a)** Cross Section **(b)** Graph

**Figure 2.5** – Transformation of Cross Section to Graph. (Source: [7])

since the vessel direction is calculated from the centreline point minus the previous centreline point. Also, limited segmentation results have been shown, with only a single simple phantom and two CT examples.

In these 2D cross-section techniques, the vessel calibre measurement is calculated as either the diameter of the maximum inscribable circle within the segmented lumen, or by using the cross-sectional area directly [69].

### 2.4.3 GroBa

One method developed with the aim of overcoming some of the shortcomings outlined above is the GroBa technique [9], developed specifically for MRA data. The key difference as compared to other techniques is the method by which the vessel calibre is estimated. A "balloon" is initialised as a single voxel on the centreline, which is grown via binary dilation until either the length is twice the width, or no more neighbouring voxels can be added. The vessel calibre is then computed from the equation of a cylinder, using the calculated balloon volume and it's half-diagonal, i.e.

$$v = \pi r^2 2h \tag{2.1}$$

$$r^2 = s^2 - h^2 \tag{2.2}$$

where $r$ is the unknown cylinder radius, $h$ is the unknown half-height and $s$ is the known half-diagonal.

The main advantage of this approach is to essentially average over a small volume of vessel, providing a more stable estimate of the calibre variation along the vessel in the presence of noise. However the method was only tested quantitatively against simple synthetic vessel models, with only a qualitative evaluation of a single WBMRA dataset presented.

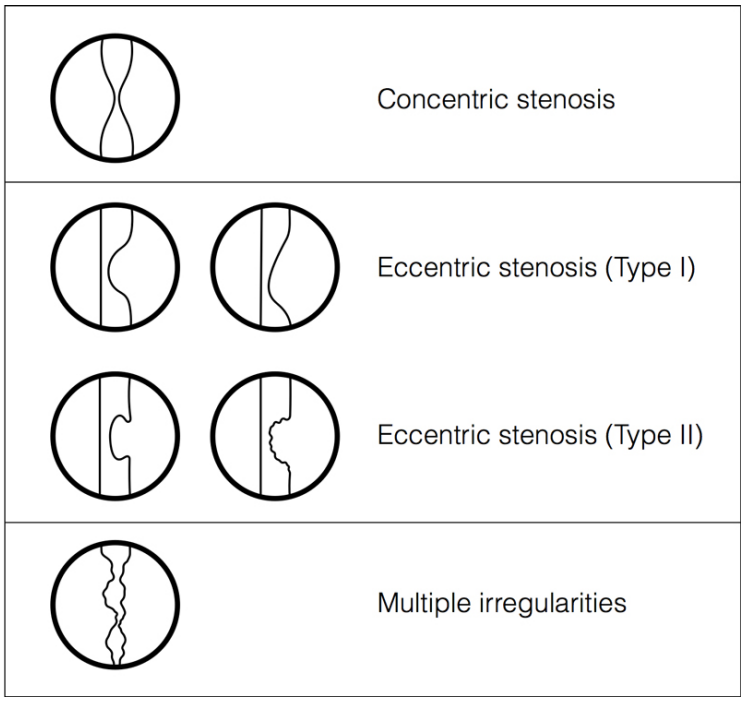



**Figure 2.6** – Types of stenosis, as described in [8].

### 2.4.4 Stenosis Detection

Stenoses can be described as having one of three morphologies as shown in Figure 2.6 — concentric, eccentric, and multiple irregularities [8]. For our data volumes all stenoses appear as concentric in all by the largest vessels (primarily the aorta), due to low effective resolution of the technique as compared to CTA.

A typical approach for detecting candidate stenoses is to look for local max-min-max in the vessel calibre measurements, thresholding to remove the candidates with small radius changes.

In [7], quantification is carried out using the average radius of the two neighbouring local maxima, $R_{left}$ and $R_{right}$, assigning a grade based on the normalised sum of the differences between the average radius of the neighbouring maxima and each radial length around the optimal boundary. This attempts to incorporate information about the variance in shape into the stenotic grade assessment, though no clinical foundation is provided for this.

Another manual approach is discussed in [69] for detecting cerebral artery stenoses. Here, vessel diameters measured using a simple cross section, rotating the vessel manually such that the centreline at the point of interest runs along the z-axis and the cross section given by the x-y plane. In the case of multiple intersections, the vessel section of interest is selected manually. A level set is the fitted to the segmented cross section

and the enclosed area calculated. Again, this approach requires too much user interaction, with the interpretation of "healthy" vessel and stenosis left to the user, so is unsuitable for a fully automated system.

In the case of CTA, where the majority of development has been focussed due to it being held as the gold standard for stenosis detection, many algorithms have been developed specifically for the coronary arteries [70]. Here, the best performing approach [53] employed a level set technique to segment the vessels [71], while using an estimated expectation of the vessel profile combined with a calcium segmentation step to calculate the vessel narrowing. It was shown that automatic lumen segmentation is possible with a precision similar to that obtained by experts, and the stenosis detection stage may be accurate enough to be used as a "second-reader" in clinical practice. A number of similar approaches have been proposed for CTA for both semi-automated [72, 73] and fully automated [18, 74, 75] vessel extraction and stenosis detection in the coronary or carotid arteries, none of which reach the sensitivity or specificity of a trained reader. Additional approaches include random forests [76, 77], spatio-temporal tracking [78], cross-section area [79, 80], and parametric intensity models [81]. In all cases these have been applied only to coronary CTA, and to our best knowledge no stenosis detection algorithms have been previously reported for WBMRA.

A number of commercial systems exist which aim to provide semi-automated tools for vascular segmentation and stenosis detection, such as the "Withinsight Automation Framework" by Claron Technology Inc. [15], "COR Analyzer" by Rcadia [16], and "VesselMASS RE" by Medis Specials [17]. Whilst there have been a number of publications relating to these tools, their use is still limited to providing a decision aid as a second reader or in some cases as a screening tool for ruling out healthy patients [18, 19].

## 2.5 Conclusions

The task of automatically tracking and measuring vessels is an area of very active research, however many newly proposed methods still often require a degree of manual input, or are localised to specific vascular regions, such as the coronary arteries. Many of the published methods of stenosis detection are also applied only to CTA due to its high resolution and use clinically as a gold standard for diagnosis. A number of commercial systems are available which offer automated or semi-automated

tools for segmenting vessels and manually measuring stenoses, though only one of these systems [15] is compatible with WBMRA. None of the systems have been shown to offer a sensitivity comparable to a trained clinician, so are typically used as "second-readers" and decision aids.

We can see WBMRA holds some unique challenges relating to detecting small, localised stenoses in a large arterial tree containing a wide range of vessel sizes and disease severities. We identified the greatest clinical need which is not currently satisfied by any published methods as being a fully automated system for identifying potential stenoses from a WBMRA volume, tested on real patients with well-characterised ground truth. To meet this aim, we first evaluated the performance of the GroBa system [9] for extracting and measuring vessels in WBMRA (Chapter 5). Taking this as our initial framework, we then identified which sections of the pipeline we would work to improve through a combination of previously published algorithms and our own methods, in order to create a system which could be applied to real patient data. As shown in Figure 1.3, this resulted in using an improved off-the-shelf registration method developed by Canon for pre-processing (Chapter 5), the collection of vessel segmentation ground truth and evaluation of existing segmentation methods against our own CNNs to improve upon GroBa's Frangi filtering [44] approach (Chapter 6), the use of the centreline algorithm described by Robert Van Uitert *et al.* [4] which calculates centrelines from the binary vessel map, developing a new method for tracking and extracting detected vessels (Chapter 8), collecting stenosis ground truth using a well defined annotation protocol (Chapter 7), and proposing a novel stenosis measurement algorithm which was tested alongside existing rule-based methods on our ground truth stenosis examples (Chapter 8).

# Chapter 3

# Clinical and Technical Motivation

## 3.1 Introduction

At the outset of the project we undertook two exercises to both understand the current clinical practice for examining WBMRA data, and to identify what the unmet demands are for diagnostic assistance software. In this chapter we first discuss the diagnostic workflow employed at Ninewells Hospital gathered from a shadowing session with one of our clinical partners during a typical reading session. We then describe a prototyping exercise that was undertaken in conjunction with our clinical partners and Canon Medical, aiming to outline a potential final application desired by clinicians and the technologies involved. This type of exercise is often an important step in commercial software development, laying the groundwork for transferring research out of the lab and into general practice, and was used here to help target our technical work towards the main areas needed to meet the clinical demands.

## 3.2 Typical Clinical Workflow

This section describes the typical diagnostic setup used at Ninewells Hospital when examining WBMRA scans for signs of cardiovascular disease. The data was collected when shadowing a typical afternoon session with a trained radiologist, during which a list of patient scans are examined in sequence.

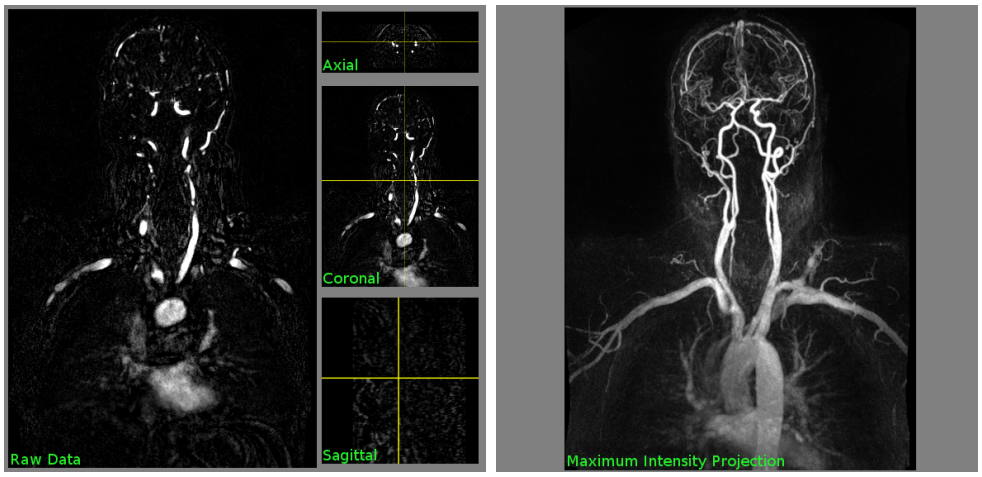



**Figure 3.1** – Mockup of general layout and contents of left & right monitors.

### 3.2.1 Computer and Software

#### Workstation

The workstation used consisted of three monitors; one standard consumer monitor was used for viewing patient records and selecting scans from the NHS databases. Two further monitors were used when examining the scans themselves. These were calibrated widescreen displays, arranged vertically, and capable of displaying high bit depth greyscale images. These were viewed in a darkened room to minimise possible reflections.

#### Software

The software used when viewing the data volumes was Carestream Vue PACS, running under Windows XP. A java-based program was used for browsing the datasets and passing them to the viewing software from linked medical records.

Figure 3.1 shows a mockup of the overall layout of the software across the two monitors. The left-hand monitor shows the raw data slices of the post-contrast dataset in the three orthogonal planes. Typically each separate station for a given patient is loaded and examined in sequence, not a single stitched volume. The post-contrast dataset is used as the additional tissues give useful positional information when viewing isolated slices.

The coronal, sagittal and axial planes are displayed simultaneously on the right of the screen, and can be dragged into the main frame to enlarge as required. The axes of the cutting planes can be moved and oriented by clicking and dragging the lines in the smaller views to the desired position/angle, with the brightness and contrast adjusted by

right-clicking the mouse and dragging (brightness mapped to the x-axis and contrast to the y-axis).

The right-hand monitor displays the coronal maximum intensity projection (MIP) of the subtracted dataset, and was used primarily for a quick initial scan to identify potential problem areas, and for reference purposes when examining the individual slices on closer detail. It also gives a quick indication of the overall quality of the scan, clearly showing any artefacts. This is also important for the final station where there are typically 3 scans taken due to variances in the timing of the contrast circulation, with only the best one being selected for examination.

### 3.2.2 Diagnostic Decision-Making

Each vessel on the main arterial tree is examined in sequence, starting on the left of the image and following the main arterial path around. The vessel paths are followed by manually scrolling backwards and forwards through the data as the vessels move in and out of the given plane. This is done by eye without the aid of any further measurement tools with the initial decisions being based purely on training and experience. The coronal/sagittal planes are oriented to show the vessel longitudinally to gauge the amount of narrowing, with the cross-section from the axial plane used for confirmation if needed. The MIP view is often only used for general location information, except in the case of the calves where the data can be of lower contrast, and the vessels are of smaller calibre. Here, the MIPs are often more useful for locating potential stenoses.

Primarily, the vessel "wall" is examined; in practice this refers to the outer surface of the lumen as the vessel wall itself generates no signal in MRA (though it can sometimes appear as a dark region around the vessel depending on the surrounding tissues and size of the vessel — see Figure 3.2). A healthy vessel exhibits a smooth, continuous surface when scrolled along the axial direction. Areas of stenosis appear as a disruption in this surface, eating into the interior of the lumen. The smooth lumen surface and slowly varying diameter of the vessel are the main features which are examined visually, with the signal strength itself giving little information due to variations in intensity from unrelated processes (blood flow fluctuations, timing errors in tracking the contrast agent circulation, etc.).

Identified stenoses are assigned a severity grade by eye; these are only deemed clinically relevant above 50% diameter reduction, where they begin to disrupt the haemodynamics within the vessel. No standardised or

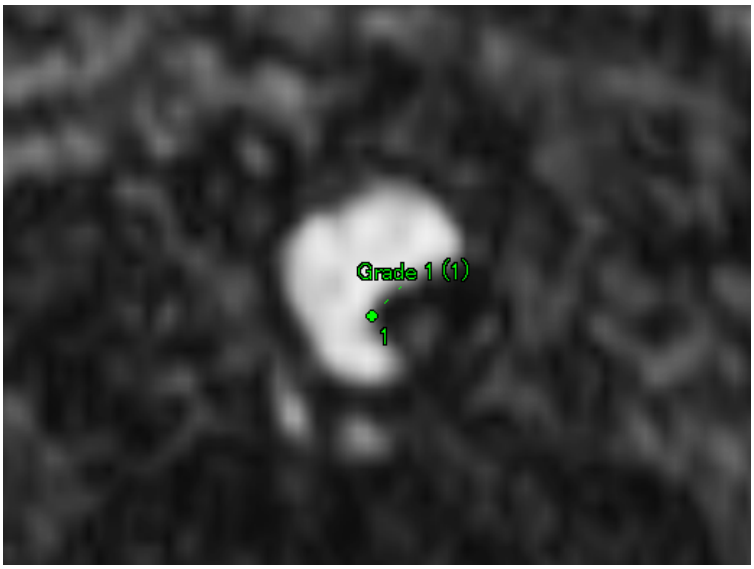



**Figure 3.2** – Example of a grade 1 stenosis in the abdominal aorta. Shows the characteristic dark vessel wall surrounding the bright lumen, with the stenosis eating into the interior of the lumen.

widely accepted grading scale exists, with different different methodologies in use at different sites, sometimes varying between different studies. The current scale used in a number of clinical studies at Ninewells ranges from 0–4. with grades cutoffs at 50% for 1, 70-80% for 2, 90-95% for 3, and 4 for a complete occlusion. Note that for the ground truth gathered for this project (Chapter 7), a scale ranging from 0–5 (Table 5.2) was chosen by clinicians due to the desire to detect stenoses before they become haemodynamically relevant, allowing them to be tracked over time and targeted earlier by existing long-term therapies.

Once detected, a stenosis grade is estimated by eye, looking at the vessel before and after the stenotic region to determine the expected width of the vessel. Occluded vessels often demonstrate many additional small neovascular paths that have formed to bypass the occlusion and ensure blood perfusion continues through the tissues, and this additional context is often used to differentiated between a very high grade stenosis (95%+) and a complete occlusion. Symmetry information from the opposite side of the body is also used for areas of complex disease or diffuse vessel irregularities.

When reporting, the final grade given is based on the most severe stenosis in that arterial segment, regardless of how many other stenoses are present. Small arteries off of the main tree are typically not examined as stenoses in these regions generally have a negligible impact on health outcomes, since the area is often well supplied by a multitude of these smaller vessels. Also there will be comparatively fewer vessels supplied

by that smaller vessel, so there is a much lesser knock-on effect further down the arterial path.

## 3.3 "Paper Prototype" of Software Tool

The aim of this prototyping exercise, undertaken with collaboration from our clinical partners, was to explore what information may be useful to be presented during the diagnosis process and how this information should be displayed. For this task we created a "paper prototype"; a mock-up interface which aims to encapsulate the main functionality of the target clinical application. This iterative process allows the interface elements to be rapidly changed and refined, with input from team members spanning a range of disciplines from Dundee University, Ninewells Hospital and Canon Medical.

Our approach was to create a mid-fidelity interface [82] in Microsoft PowerPoint, which allowed interactivity with the main functions of the envisioned program. We aimed to identify all the image analysis and visualisation technology required to meet this end goal. Our target was deliberately ambitious when formulating the target application with clinicians, incorporating as many broad features and functionalities as they felt were needed regardless of whether or not they were feasible within the scope of this project alone.

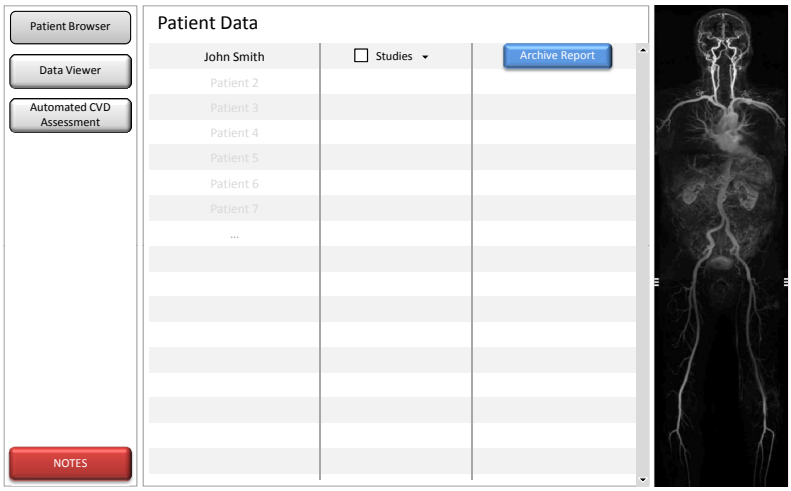
### 3.3.1 Key Features

Figure 3.3 shows example screens from the three main portions of the developed interface; the patient database browser, MRA data viewer, and automated CVD assessment window.

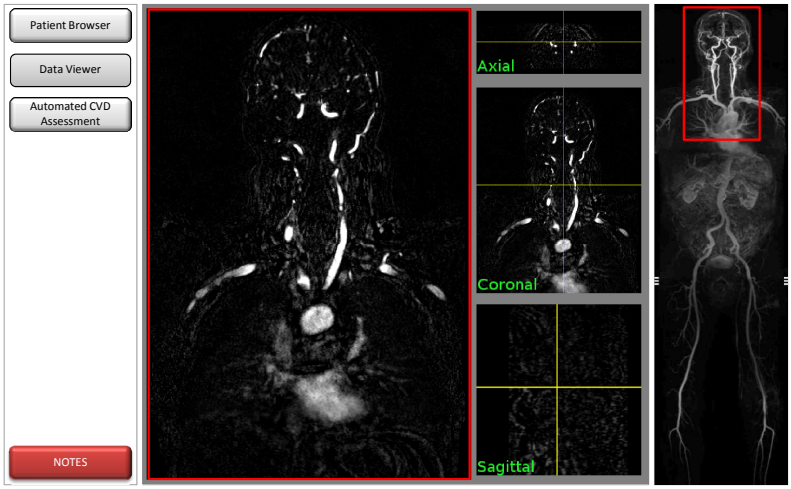
Based on our discussions with the clinicians, the following desirable features and interactive functionalities were identified.

#### Data Inputs for Processing and/or Visualisation

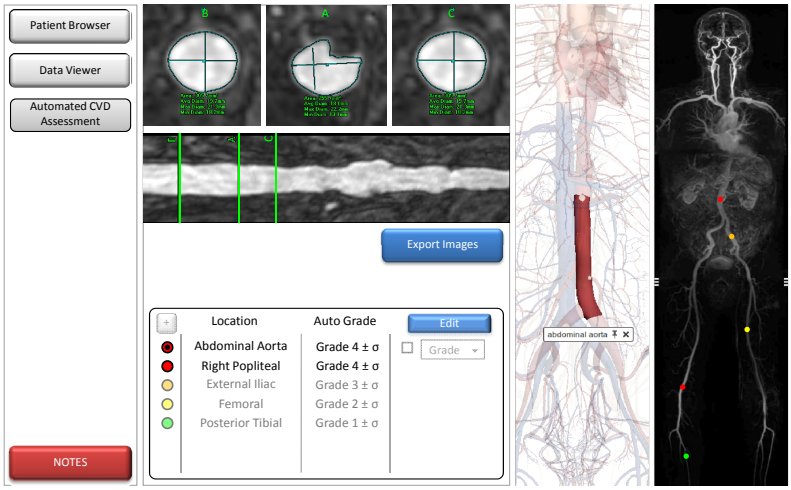
1. Importing of MRA datasets; whole body (as separate stations), or region of interest (e.g. run-off studies). 2D slice and 2D MIP views desired. 3D volume rendering believed to be of limited diagnostic use in the current clinical workflow.
2. Functionality to view other relevant tests such as blood tests, kidney time-activity curves, etc., in conjunction with the imaging data.
3. Visualise functional data for other regions and/or organs, e.g. cardiac assessment (perfusion, volumes), in addition to relevant studies




**(a)** Data Importer




**(b)** Volume Slice Viewer




**(c)** Automatic Stenosis Assessment

**Figure 3.3** – Example screens from the interface prototype of a proposed WBMRA vascular analysis program, showing the patient database browser, MRA data viewer, and the automated CVD assessment.

from other imaging modalities, e.g. X-ray CT, carotid vessel wall thickness measurements from ultrasound, etc.

**Display, Functionalities and Interactions**

1. Automated WBMRA lumen assessment, with graded stenoses. The areas examined, and measurements used, should be presented in an easily readable way for the clinician to confirm.
2. Function to allow the confirmation/alteration/addition of stenosis scores by clinician, presented as an ordered list with drop-down menus.
3. Use of the WBMRA MIP image as an interactive navigational tool, for displaying location of stenoses and possibly pulling up additional data relevant to the overall diagnosis, e.g. click/tap on kidney to pull up kidney perfusion data.
4. Flexible interface to allow arrangement of data viewer windows as needed, possibly with different modalities, e.g. viewing MRA and CTA side-by-side. Current solutions which are available tend to result in user overload, with too many windows to organise, and are therefore not often used unless absolutely necessary.
5. Labelling of each stenosis grade on whole body MIP as a different coloured spot, with the ability to toggle on and off different grades, and switch between displaying all stenoses and just the worst stenosis per artery.
6. Possibility to integrate 3D vessel atlas (e.g. Zygote body model) as a potential teaching/training aid.
7. Automatically orient data viewer to given stenosis of interest, allowing the clinician to quickly validate the assigned grade without having to manually locate it in all three axes.
8. 3D volume renders of limited vessels of interest may be useful aid for surgery planning.

**Outputs**

1. Automated whole-body CVD assessment, presented as a list of stenosis grades per artery with confidence levels, verified by clinician.
2. Clinical flags for problem areas which could be tracked over time, for assessing disease progression.
3. Generate a standardised report, populated by the results of the program, which can be archived and viewed at a later time.

4. Archival of automatically generated images of problematic areas (cross-sections and MIPs) for easy display and discussion at Multidisciplinary Team Meeting (MDT).

### 3.3.2 Key Technologies

In order to meet the demands of the clinical tool outlined above, a number of key technologies were identified. These could be grouped into two categories; those requiring additional research to realise, and those which relate to software development.

#### Software Development

Should a commercial software package be developed for assisting CVD diagnosis from WBMRA data, the following functionality was highlighted as being desirable:

1. Data viewer (similar to Canon Medical's current Voxar3D viewer) to display slices through the volume with associated interactive tools (window level adjustment, etc.). Keeping as close as possible to the current software in use would ease training and encourage adoption.
2. Display of Curved Planar Reformation (CPR) views for visualisation of curved vessels in 2D.
3. Possible 3D rendering of segmented lumen for surgery planning.
4. Integration with other databases to access and display other relevant tests to aid with the diagnosis (e.g. blood tests, X-Ray CT etc.).

#### Areas of Further Research

The following areas were identified as needing further research in order to realise the outlined prototype system:

1. An automatic vessel segmentation algorithm, optimised and validated using WBMRA data.
2. An automated stenosis detection algorithm leveraging either local vessel calibre measurement algorithms or appropriate machine learning techniques, validated against ground truth stenosis measurements.

3. A vessel landmark atlas for localising and naming vessels, enabling generated CVD assessment reports with named vessel segments, integration with teaching aids (such as Zygote model), and potential improvements to 3D vessel tracking.

Of these, the first two were determined to be the most important to realise the core functionality of the system (automated stenosis detection and grading), with the third providing additional features and potential improvements to existing solutions for vessel tracking.

## 3.4 Conclusions

In this chapter we have outlined the workflow of a typical WBMRA diagnosis session through a shadowing session with a trained radiologist, examining what decisions are made during the process and how these are done. Using that information, alongside input from our clinical partners, we developed a mid-fidelity paper prototype interface in Microsoft Powerpoint of a diagnostic aid which would meet the clinical demands of this task. Through this exercise of prototyping a user interface for clinicians we identified a number of visualisation and image analysis technologies which must be developed and/or refined in order to meet the goal of aiding the diagnosis and grading of CVD from WBMRA examinations.

One of the most important aspects of such a diagnostic aid, emphasised strongly during our meetings with radiologists, is the need for a fully automated tool. Manual and semi-automatic measurement tools already exist, but these are rarely if ever used in practice due to the time pressures of a clinical environment and additional training required, and are often considered to give limited benefits over estimation by eye.

The key aim for a new diagnostic aid should therefore be to develop a fully automated system which does not require any user interaction, and which presents the results to clinicians in an intuitive way, allowing interaction as needed for modifying or confirming the measured regions of disease. It is also important that the underlying algorithms have been throughly validated with well-characterised ground truth data so that clinicians can be confident in the results presented.

The paper prototyping exercise highlighted a number of areas which require further research, as detailed in Section 3.3.2. In this thesis we target the two key research areas which arose; the need for an automatic vessel segmentation algorithm optimised and validated using WBMRA data, and an automated stenosis detection algorithm validated against

ground truth stenosis measurements. To meet these challenges we gathered ground truth segmentation maps (Chapter 6) and ground truth stenosis annotations (Chapter 7), compared a number of common segmentation techniques alongside bespoke convolutional neural networks trained for WBMRA vessel segmentation (Chapter 6), and developed and tested three automated stenosis detection algorithms for assigning stenosis grades to segmented vessels (Chapter 8).

# Chapter 4

# Materials and Preprocessing Methods

## 4.1 Introduction

Three datasets were used in the experiments presented in this thesis; one set of raw WBMRA patient scans, and two sets of ground truth data for technique development and performance analysis. No public datasets for WBMRA currently exist, so all data used was acquired specifically for this study or a companion project. This chapter will briefly describe these datasets, how they were acquired, and what they were used for. A more thorough description of the protocols used, and analysis of the data quality and variability, will be provided in their respective chapters as indicated.

## 4.2 WBMRA Datasets

For this study a total of 18 patient WBMRA datasets were used; 10 male, 8 female. These were previously acquired as part of the TASCFORCE project [83] (ethics approval Tayside Committee on Medical Research Ethics B, 13/07/2007, ref: 07/S1402/42) using a 3.0 Tesla MRI scanner (Siemens Magnetom Trio), and acquired as 4 separate image volumes, or "stations", per patient. Station one comprises the head and neck, station 2 the thorax and abdomen, station 3 the pelvis and thighs, and station 4 the feet (see Figure 4.1). In addition, both a "pre-contrast" and "post-contrast" volume were provided, whereby a gadolinium-based contrast agent was injected prior to the acquisition of the second volume, generating high contrast within the arterial lumen.

The datasets themselves were provided as a sequence of coronal slices

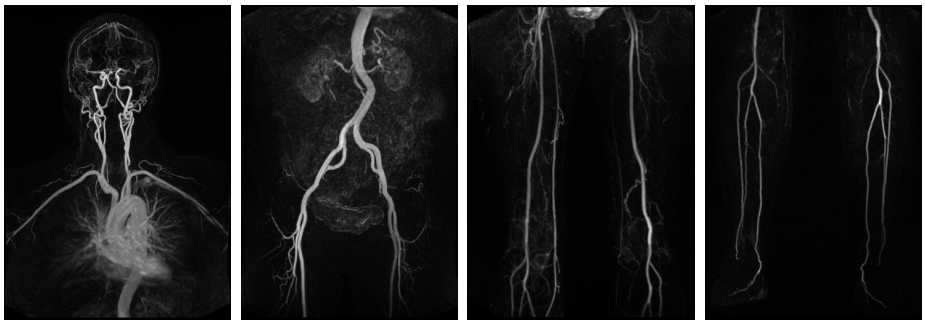


**Figure 4.1** – Maximum intensity projections of the four stations of a typical patient, shown after digital subtraction of the pre-contrast from the post-contrast volumes.

for each of the 4 stations, in the standard DICOM format. These had a resolution in the coronal plane of $0.98mm \times 0.98mm$, with a slice spacing of between $0.98mm$ – $1.3mm$.

The patient cohort consisted of 14 symptomatic peripheral arterial disease (PAD) patients and 4 healthy volunteers. Patient datasets were acquired as part of a previous WBMRA study on symptomatic PAD [84], from which 18 were selected to provide a range of atheroma burden severities according to a previously assigned standardised atheroma score (SAS) [85]. All of the selected PAD patients had evidence of atherosclerotic plaques at multiple sites. Healthy volunteers had no previous history of atherosclerotic disease and matched by age and gender frequencies to the PAD patients. As the data was acquired for a previous study and supplied in a fully anonymised state, information such as the age, gender, etc. of each specific patient was not made available. The authority for how the data could be handled also lay with the principle investigator of the original study, and authorisation for releasing the datasets publicly was not granted.

## 4.3 Ground Truth Vessel Segmentation

Segmentation ground truth was acquired as part of our project. A manual segmentation protocol was created and applied to three of the patient datasets, as described in Chapter 6.

The binary vessel segmentation volumes were saved as tiff stacks with a separate file for each station, mirroring the structure of the original data. Figure 4.2 shows the maximum intensity projections of these three datasets, which were used to optimise and validate a range of segmentation algorithms as described in Chapter 6.

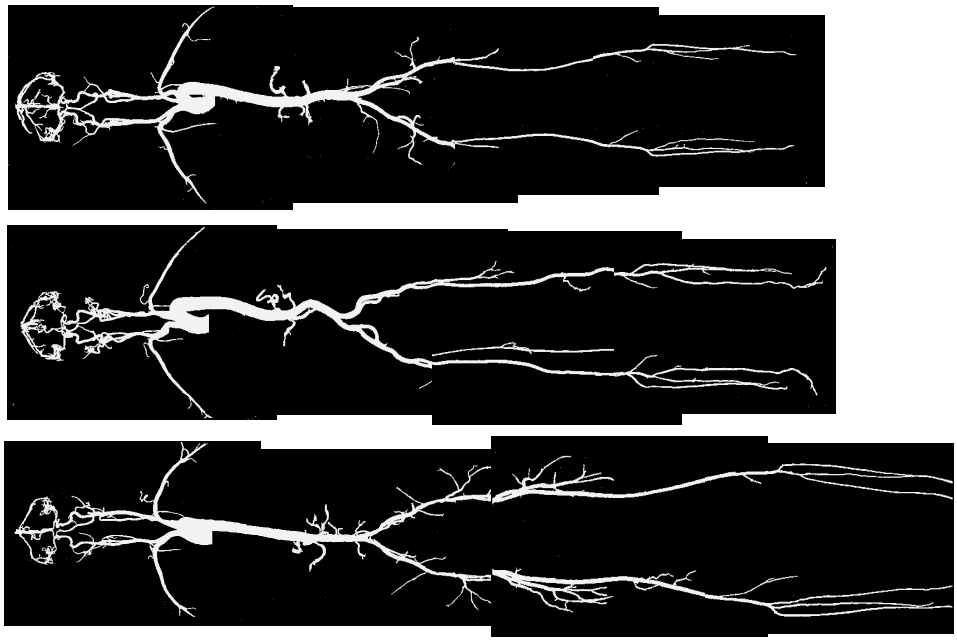


**Figure 4.2** – Maximum intensity projections of the three ground truth segmentation sets created for this project.

## 4.4 Ground Truth Stenosis Assessments

The second ground truth set acquired for this project consisted of stenosis severity assessments for our 18 patients. This was done by recruiting 3 trained clinicians to examine 37 pre-defined arteries, assigning them a score between 0 – 5 based on the maximum reduction in lumen diameter (Figure 4.3). The developed annotation protocol is discussed in Chapter 7, along with a detailed analysis of the variability discovered between the three trained annotators, each of whom had different clinical backgrounds (research, diagnosis, and intervention) which lead to a high variability in the lowest severity grade in particular. We also discuss the implications this has for any system using such data as ground truth.

## 4.5 Data Preprocessing

Three preprocessing steps were applied to the raw patient volumes:

**Data Interpolation -** The patient volumes had a variable slice thickness of between $0.98mm - 1.3mm$.We therefore first applied a linear interpolation in the slice direction to give an isotropic voxel dimension of $0.98mm \times 0.98mm \times 0.98mm$. This allowed us to track

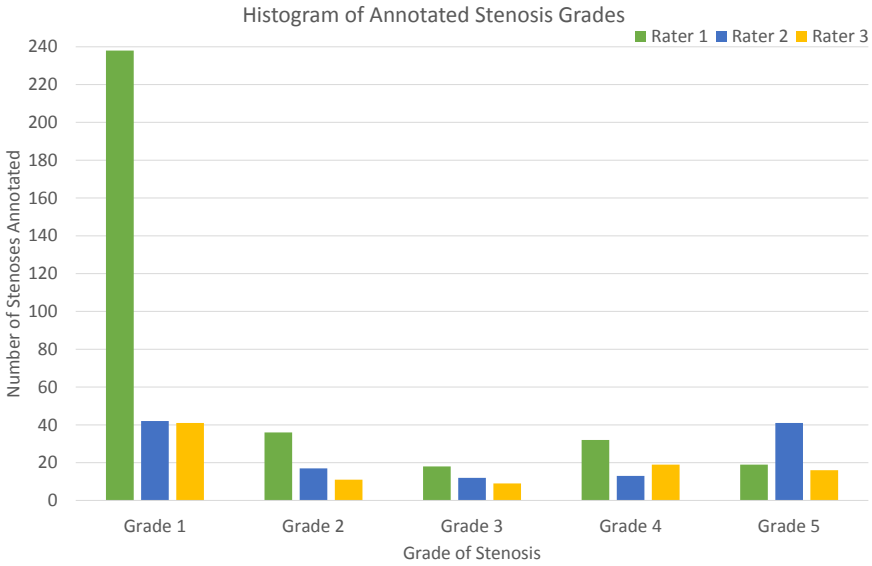



**Figure 4.3** – Total number of stenoses found per grade by each rater, for all 18 patients.

and measure vessels in 3D without any geometric distortions due to anisotropic voxels.

**Volume Registration and Subtraction -** Our data consisted of two scans, a pre- and post-contrast scan, as described in Section 4.2. Due to the significant scan times, and time between scans, volume registration had to be applied to correct for any patient motion. The registered pre-contrast volume was then subtracted from the post-contrast volume to suppress static tissues. This process is described in Chapter 5.2.

**Voxel Masking -** In MRA datasets, slices at the extreme edges of the volume can typically show a number of artefacts (low signal-to-noise ratio, intensity artefacts due to phase wrapping, etc.), as well as being no longer overlapped by the companion scan due to the registration process. A final artefact correction step was therefore applied, masking border voxels to remove MR artefacts and tissues not covered by the pre-contrast volume.

## 4.6 Conclusions

In this chapter we have described our patient dataset, and the ground truth datasets used in the validation and testing of our WBMRA processing pipeline. No public WBMRA datasets currently exist, therefore all three datasets used in our study were acquired by ourselves and our clin-

ical partners and are not public. A thorough description of the protocols used, and analysis of the data characteristics, are provided in Chapters 5, 6, and 7. In the next chapter, we will discuss an evaluation of the baseline methods which our work builds upon.

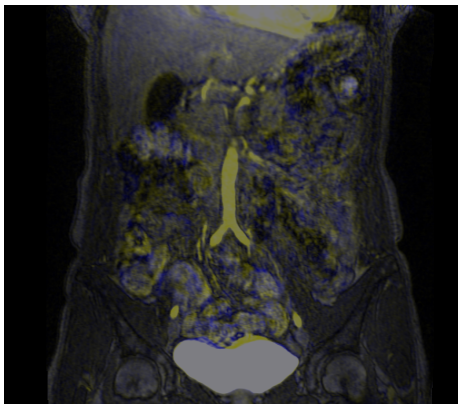
# Chapter 5

# Data Pre-Processing and Baseline Method Evaluation

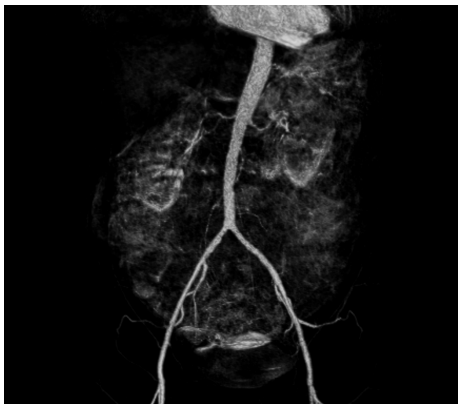
## 5.1 Introduction

This chapter will discuss the first experiments undertaken at the beginning of this project. As described previously, the algorithms in our pipeline are designed to work with digitally subtracted volumes whereby the pre-contrast volume is subtracted from the post-contrast volume, suppressing static tissues and leaving only information about where the injected contrast agent has travelled. Since there is a significant time delay between the two volume acquisitions, a volume registration step must be carried out in order to minimise artefacts. Therefore, during an initial industrial placement at Canon an evaluation of two registration software packages was performed in order to determine which should be used for this pre-processing stage. The resulting subtracted data is common to all further experiments in this thesis and will thus be discussed here first in Section 5.2.

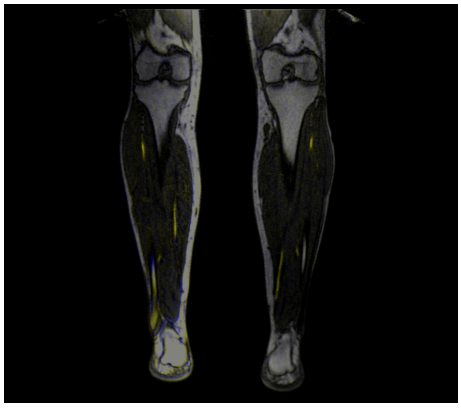
Secondly in this chapter we examine the only previously published method for automated lumen calibre measurement in WBMRA, set out by Perez-Rovira *et al.* in [9]. Taking this method as our baseline, we evaluate its performance in a range of synthetic tests (Section 5.3.2) in order to determine its strengths and shortcomings.



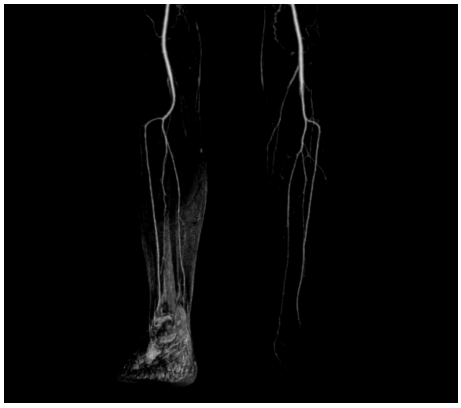
**(a)** Coronal view of without registration (Station 2)



**(b)** MIP after Siemens registration applied



**(c)** Coronal view of without registration (Station 4)



**(d)** MIP after Siemens registration applied

**Figure 5.1** – (a) and (c) show the coronal views of the fused pre- and post-contrast images, with (b) and (d) showing the coronal MIPs of the corresponding section in the subtracted WBMRA data using the Siemens software. Visualisations were created using Canon's RWS software, with blue regions in (a) showing areas of higher intensity in the pre-contrast scan and yellow regions showing higher contrast in the post-contrast scan.

## 5.2 Data Pre-Processing: Volume Registration

When combining medical images acquired at different times errors are typically introduced due to patient movement between the two acquisitions. This can be a significant issue for WBMRA scans, which can typically take around an hour, with errors caused by accidental motion — such as the patient relaxing their shoulders or moving their leg — or unconscious processes such as peristalsis waves in the intestine.

The patient volumes initially provided by Ninewells Hospital had been processed using a commercial software package provided by the MRI scanner manufacturer (Siemens). This automatically registers the pre-contrast volume to the post-contrast volume, subtracts them, and then

stitches the four stations into a single large volume. No information is available from Siemens as to the exact algorithms used in their commercial software, however based on the appearance of certain artefacts and artificial markers inserted into the processed images we believe it to be based on the method presented by Wachinger *et al.* in [86].

Figure 5.1 shows an example from stations 2 (abdomen) and 4 (feet) which exhibit noticeable artefacts. In the left hand images the pre-contrast dataset has been coloured blue and the post-contrast dataset coloured yellow, and they are overlaid without any registration applied. Where the data overlaps correctly, the blended result is greyscale. In this way, contrast in the arteries shows as bright yellow, and bulk movements can be visualised as a mirrored blue and yellow halo in the direction of motion. The right hand images show a MIP of the same station after registration by the Siemens software.

We can see that the current processing software used does not always correct for movement accurately; in Figure 5.1a internal motion in the abdominal cavity and expansion of the bladder leaves residual signal from the organs in the subtracted dataset, and in Figure 5.1c the patients left leg moved slightly between acquisitions, leaving a "shell" around the limb in the subtracted view.

Volume registration for medical images can be a highly complex task, and has been an area of extensive research for decades [87–91]. In consultation with the clinical team and Ninewells and industry partners at Canon it was concluded that the since the registration results were good enough to be used for human diagnosis then any errors were unlikely to become a limiting factor in our future research, and therefore developing a bespoke solution should not be an aim of this project. However, it was recognised that a more accurate registration of the pre-contrast volume to the post-contrast volume may help reduce the artefacts left after the subtraction stage, resulting in a cleaner reconstruction. Canon therefore provided its own internally developed registration program, named "tmvsFire", for comparison. This was a pre-compiled tool for Windows, run from the command line. No access was given to the source code, so only the information provided in the company-confidential user report is detailed here.

This section will describe the qualitative tests carried out on our WBMRA datasets to compare the performance of these two registration solutions. Since no registration ground truth exists for our data, the main focus of the study was to determine which of the two available solutions

minimised artefacts in the subtracted volumes.

### 5.2.1 Materials and Methods

#### Siemens Registration Software

The clinical software in use at Ninewells hospital was provided with the scanner by its manufacturer, Siemens. It is therefore treated as a black-box since no information is provided as to the methods or parameters used.

#### Canon's "tmvsFire" Software

tmvsFire is a general purpose Windows command line tool created by Canon Medical, implementing two registration algorithms; a global, "rigid" algorithm, correcting for translation, rotation and/or scaling, and a local, "non-rigid" algorithm for refining individual voxel positions. These can be run either singly, or both in combination.

The registration framework is based on the method of Crum *et al.* [92], which uses mutual information (MI) [93,94] as both the global similarity metric and to calculate the forces used to drive a fluid deformation model, and is combined with the fluid/elastic regularisation framework based on Gaussian smoothing described by Thirion [95]. This information was provided by the author of the tool in the user documentation [96], but no access was given to the source code so this information is quoted here verbatim and without further extrapolation as the exact implementation and parameters were not provided.

The tool itself reads regular and enhanced DICOM files, writing the resulting data as a series of regular format DICOM images. A registration mask, provided as a bitmap (DICOM, MetaImage, ANALYZE or PGM formats), can also be used to label voxels that should be ignored during the registration calculations, such as where motion outside the area of interest negatively affects the registration of important regions, however this was found to be unnecessary for our data.

#### Datasets

As described in Section 4.2, the datasets to be processed consisted of 18 WBMRA datasets consisting of a pre- and post-contrast scan for each of the 4 stations (head, abdomen, legs, feet). An additional subtracted and stitched whole-body dataset was created using the standard Siemens re-

construction software in use at Ninewells Hospital, to which the tmvsFire results were compared.

#### Additional Software

MATLAB R2013b was used for creating the subtracted datasets, and ImageJ 1.47v used for manual inspection of the registered volumes.

### 5.2.2 Registration Results

#### Data Preparation

The DICOM datasets used consist of a sequence of coronal slices for each of the 4 stations, as previously described. When these were initially supplied, it was found that they were anisotropic, having voxel dimensions of $0.98mm \times 0.98mm \times 0.98 - 1.3mm$. As it is preferable to work with isotropic data for vessel tracking and measurements in 3D, both the pre- and post-contrast datasets were interpolated to a voxel dimension of $0.98mm \times 0.98mm \times 0.98mm$ before being passed to tmvsFire. The files processed by the Siemens software remained anisotropic, so were interpolated using the same algorithm in order to be directly comparable with the tmvsFire results.

#### tmvsFire Parameters

tmvsFire has a number of parameters that can be altered to change the behaviour of the registration algorithm. Each parameter also has a built-in default value, which been previously set empirically by the developer after testing the algorithm across a range of proprietary X-ray CT and MRI datasets in Canon's databases during its development [96].

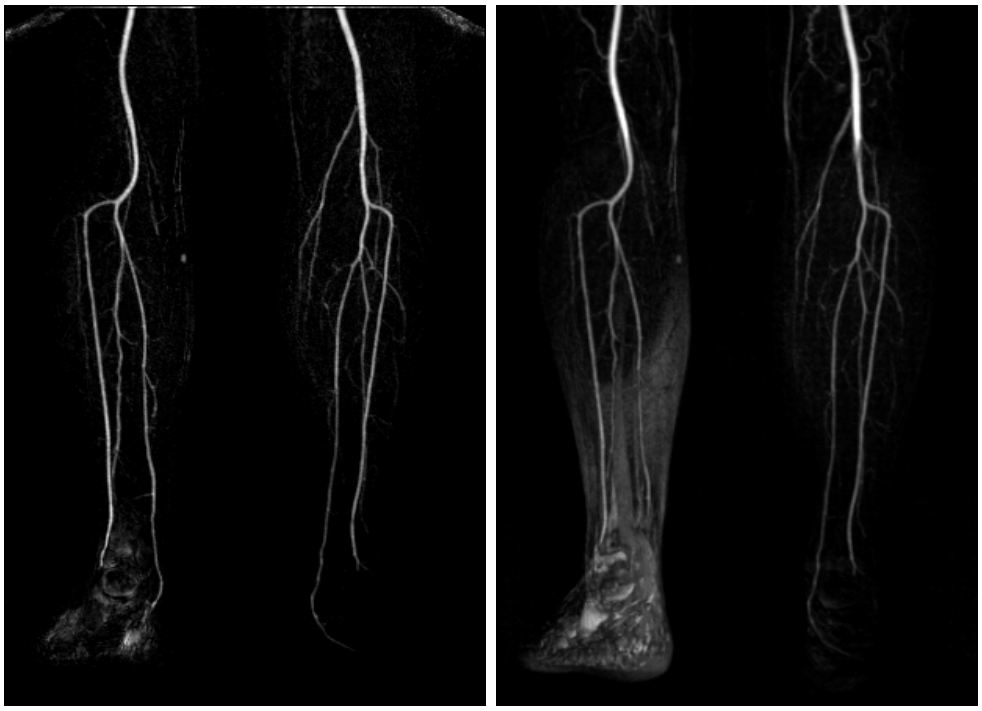
Listed below are the parameters that could be set at runtime, along with the default values (shown in square braces):

#### Histogram Binning

**`-k`** — Use k-means binning [0]

#### Rigid

**`-R`** — Rigid registration [1]
**`-T`** — Translation limit [0]
**`-O`** — Rotation limit [30]
**`-M`** — Rigid metric [MI]
**`-SR`** — Multiscale (rigid) [(4,2,1)]



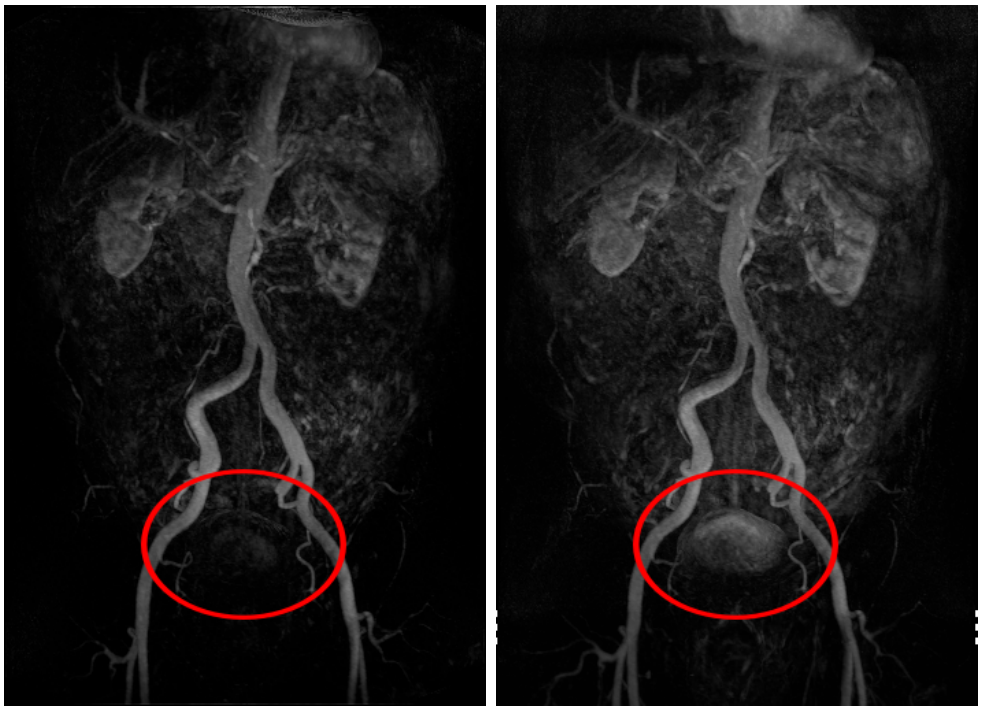
**(a)** tmvsFire **(b)** Siemens

**Figure 5.2** – Comparison between tmvsFire and Siemens registration on a patient with significant leg movement between the pre- and post-contrast scans.

**Non-Rigid**

- **`-A`** — Fluid kernel scale-factor [0.95]
- **`-B`** — Fluid kernel initial SD [19]
- **`-b`** — Forcefield engine [MI]
- **`-C`** — Constrain (zero) force field in x,y,z axis [0]
- **`-D`** — Maximum displacement/iteration [1]
- **`-d`** — Maximum MI decrease stop criteria [-0.05]
- **`-E`** — Apply elastic constraint [1]
- **`-L`** — Average improvement/step limit [-0.05]
- **`-N`** — Maximum number of iterations [40]
- **`-S`** — Multiscale (non-rigid) [(8,4,2,1)]
- **`-V`** — Apply viscous constraint [1]
- **`-Y`** — Elastic constraint kernel SD [0.05]
- **`-Z`** — Fluid constraint minimum SD [5]

Figures 5.2 and 5.3 show example results of applying tmvsFire with the above parameters, registering the pre-contrast to the post-contrast volumes followed by subtraction of the pre-contrast volume from the post-contrast volume. Visually, the results of the two programs were found to be largely comparable across most patients with the main differences occurring around areas of high motion. Figure 5.2 shows that tmvsFire



**(a)** tmvsFire **(b)** Siemens

**Figure 5.3** – Initial results of registering the post-contrast dataset using default parameters. Subtraction was done using ImageJ, with the levels adjusted by eye to match those of the stitched WBMRA dataset from the Siemens software. The bladder (circled in red) is more accurately registered by tmvsFire using the default settings.

provides a marked improvement around the region of movement in the legs, reducing the residual artefacts and increasing vessel visibility. Similarly, an improvement can be seen around the bladder in Figure 5.3. The residual artefacts from the motion of the intestine are also slightly improved in the tmvsFire image, though the kidneys appear similar.

A number of tests were carried out to attempt to optimise the tmvsFire parameters for our particular datasets, however these proved inconclusive in the absence of ground truth and attempts to estimate the quality of registration through metrics such as the number of connected components in the subtracted volume after applying a threshold did not offer consistent results across different patients.

Therefore, in the absence of quantitative data showing improvement across all patients for a given parameter value, the parameters were left at their recommended defaults. The decision on which algorithm to choose to pre-process all of our data with therefore focused on a visual inspection of three key areas which could potentially impact the performance of our system:

1. Which algorithm leaves the least obtrusive artefacts due to motion?

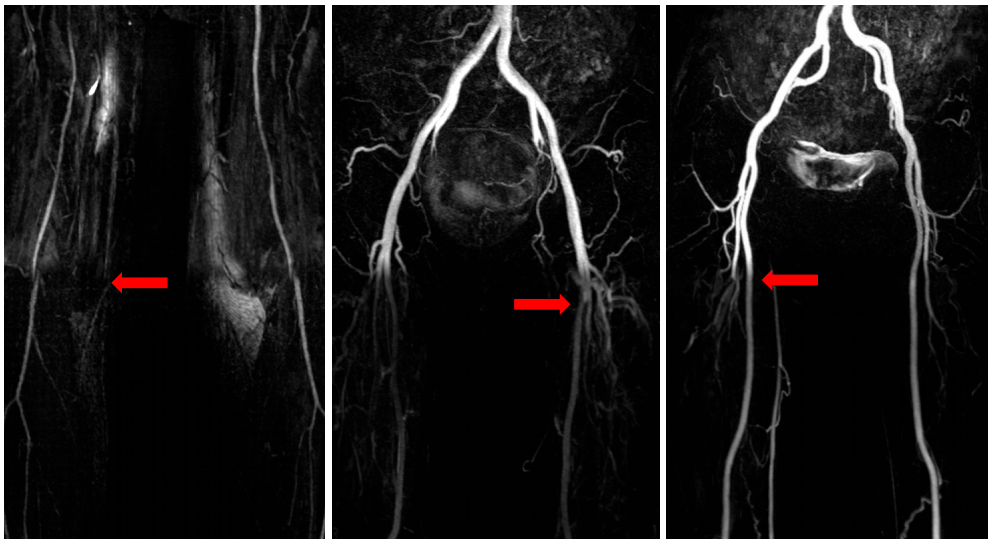


**Figure 5.4** – MIPs of three stitched datasets, with the red arrows highlighting stitching artefacts left in the overlap regions by the Siemens software.

2. Does either algorithm introduce artefacts that affect the appearance or contrast of the vessels themselves?
3. Is there a tangible benefit to using unstitched vs. stitched data?

Figures 5.2 and 5.3 demonstrate that the tmvsFire tool was able to deal with motion artefacts better, and leave fewer non-arterial residual signals in the subtracted volumes. A visual inspection of the arteries in the processed volumes revealed that neither algorithm appeared to introduce artefacts from registration that affected the arterial contrast.

However, the stitching algorithm used by the Siemens software did introduce visible artefacts in the overlap regions. In order to connect vessels between different stations they often have to be warped in the overlap regions to prevent discontinuities (this can be noted in Figure 5.2, where the two arteries at the top of each image enter at slightly different angles). In the cases where this fails, the misaligned vessels could introduce an artificial narrowing or occlusion of the vessel. In addition, the software does not apply any normalisation to the vessel intensities between stations, often resulting in sudden changes in vessel contrast between stitched stations. Figure 5.4 highlights examples of these two vessel artefacts.

### 5.2.3 Registration Conclusions

During an industrial placement at Canon a limited study was undertaken to determine which of two volume registration programs should be used

to pre-process all of our patient data volumes, the tmvsFire tool developed internally by Canon, or the commercial software used at Ninewells hospital, supplied by Siemens. In additional to volume registration and subtraction, the Siemens software incorporated a stitching algorithm to combine all four stations into a single larger volume.

From the results of applying tmvsFire and the Siemens software to our patient datasets we discovered that both achieved visually similar results, with the largest differences being found in motion artefacts in the abdominal region (station 2) and movement of the extremities (station 4). It was also found that the stitching algorithm of the Siemens software could distort vessels in the overlap regions and produce sharp intensity changes between data from different stations. For these reasons it was decided that all subsequent work should be carried out on unstitched data processed by the tmvsFire program in order to minimise artefacts in the subtracted volumes. Registration was applied using tmvsFire with the parameter values quoted in Section 5.2.2, and the resulting subtracted volumes used for all experiments described throughout the rest of this thesis.

## 5.3 Baseline Lumen Calibre Measurement System: "GroBa"

Developed at Dundee University, "GroBa" is an automated vessel analysis and measurement system written in MATLAB [9]. It incorporates a standard segmentation and vessel tracking pipeline, followed by a 3D vessel calibre measurement approach based on growing "balloons" within the segmented vessels, and was developed and validated on synthetic, idealised vessels and a single WBMRA dataset without ground truth.

As the only system currently developed specifically for whole-body MRA analysis, this approach was taken as the baseline method from which we would build. In this chapter we assess the performance of the system on both synthetic vessels and real data, and discuss a number of further tests which were carried out to determine potential sources of error in the approach.

### 5.3.1 GroBa Overview

GroBa is a lumen calibre measurement technique, based on growing balloons within a segmented vessel. The aim of the technique was to develop

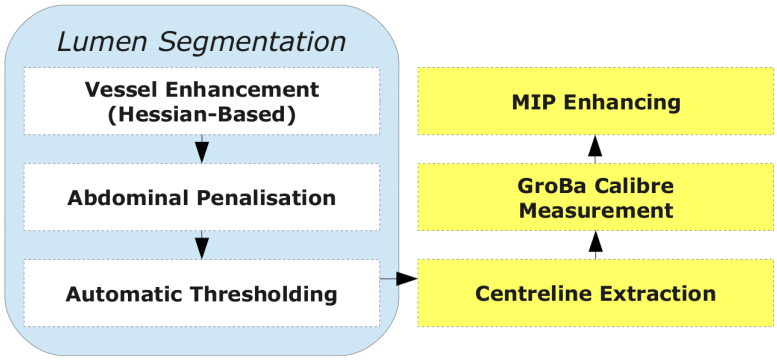



**Figure 5.5** – Schematic overview of GroBa system.

a method which could more easily cope with irregular, non-tubular vessel structures such as stenoses or aneurysms as compared to previous methods, while still obtaining accurate calibre measurements even in the presence of noisy or otherwise inaccurate centrelines.

Figure 5.5 shows a general overview of the stages of the automated system developed in [9], which segments the vasculature, obtains the corresponding centrelines, measures the lumen calibre throughout each segmented vessel and finally presents the calibre information as an HSV (Hue, Saturation, Value) colour-space overlay on the MIP images.

In this section we will describe the methods used for each of these stages.

### Vessel Segmentation

The first stage of the pipeline is segmentation of the vessel lumen, involving enhancement of vessel-like structures through an intermediate vesselness map, followed by artefact reduction in the form of an abdominal penalisation function and automatic thresholding to generate the final binary segmentation map.

The method of vessel enhancement chosen was the Hessian analysis based method developed by Alejandro Frangi [44]. This technique looks at the local curvature of the image by analysing the Hessian (local second derivative) [44, 45], extracting the principal directions in which the local second order structure of the image can be decomposed through eigenvalue analysis. From this, the direction of smallest curvature (i.e. the direction along the vessel) can be determined.

To derive the vesselness function $\lambda_k$ is first defined as being the eigenvalue with the k-th smallest magnitude, i.e. $(|\lambda_1| \leq |\lambda_2| \leq |\lambda_3|)$. Therefore, for an ideal tubular structure in a 3D image:

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | **Orientation Pattern** |
|---|---|---|---|
| N | N | N | Noisy, No Preferred Direction |
| L | L | H− | Plate-like Structure (Bright) |
| L | L | H+ | Plate-like Structure (Dark) |
| L | H− | H− | Tubular Structure (Bright) |
| L | H+ | H+ | Tubular Structure (Dark) |
| H− | H− | H− | Blob-like Structure (Bright) |
| H+ | H+ | H+ | Blob-like Structure (Dark) |

**Table 5.1** – Possible patterns, governed by the value of the $\lambda_k$ eigenvalues (H=High, L=Low, N=Noisy (usually small), +/- indicate the sign of the eigenvalue).

$$\begin{aligned} |\lambda_1| &\approx 0 \\ |\lambda_1| &\ll |\lambda_2| \\ |\lambda_2| &\approx |\lambda_3| \end{aligned} \tag{5.1}$$

In other words, the curvature should be large in the two directions (around the circumference of the vessel), and very small along the length of the vessel. Table 5.1 shows the interpretation of the local structure given the value of curvature in the three principle directions, with a visual representation given in Figure 5.6.

In order to differentiate between blob- and plate-like structures, two dissimilarity measures were proposed [44]

$$R_A = \frac{(\text{Largest cross sectional area})/\pi}{(\text{Largest axis semi-length})^2} = \frac{|\lambda_2|}{|\lambda_3|} \tag{5.2}$$

where $R_A \in [0,1] \subset \Re$ is the deviation from a plate-like structure, and

$$R_B = \frac{Volume/(4\pi/3)}{(\text{Largest cross sectional area}/\pi)^{(3/2)}} = \frac{|\lambda_1|}{\sqrt{|\lambda_2\lambda_3|}} \tag{5.3}$$

where $R_B \in [0,1] \subset \Re$ is the deviation from a blob-like structure.

In order to distinguish geometric structures from simply background noise a further "second-order structureness" measure was also defined as

$$S = ||H||_F = \sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2} \tag{5.4}$$

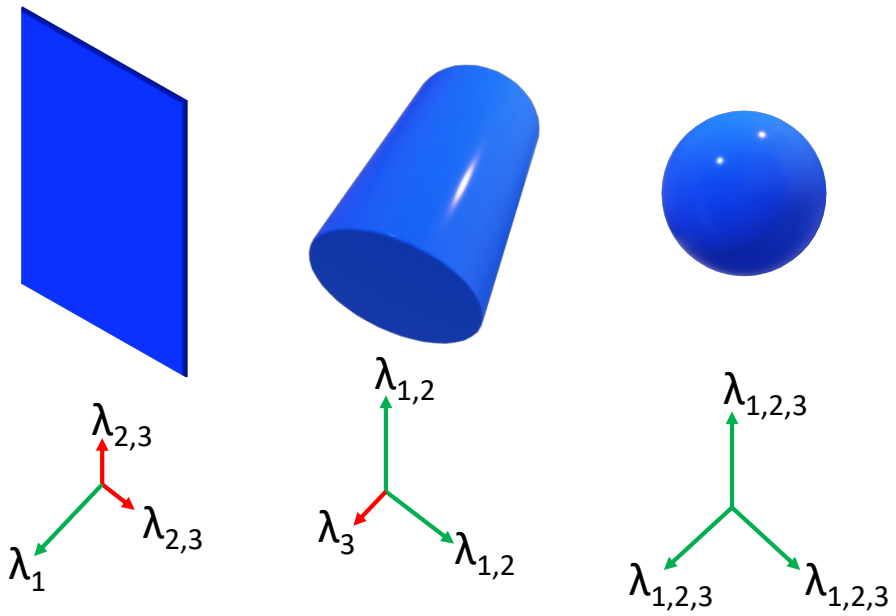



**Figure 5.6** – Isosurface representations of plate-like, line-like, and blob-like structure, according to Table 5.1.

where $||H||_F$ is the Frobenius matrix norm of the Hessian.

Combining these measures, a vesselness function was derived as being [44]

$$V_0(s) = \begin{cases} 0 & \text{if } \lambda_2 > 0 \text{ or } \lambda_3 > 0 \\ \left(1 - \exp\left(-\frac{R_A^2}{2\alpha^2}\right)\right) \exp\left(-\frac{R_B^2}{2\beta^2}\right) \left(1 - \exp\left(-\frac{S^2}{2c^2}\right)\right) & \text{otherwise} \end{cases} \tag{5.5}$$

where $\alpha$, $\beta$ and $c$ are thresholds which control the sensitivity of the line filter to the measures $R_A$ $R_B$ and $S$.

For the application of enhancing vessels in 3D angiographic datasets, the vesselness measure in Equation (5.5) is analysed at different scales, $s$. It logically follows that the response of the line filter will be maximum at a scale which approximately matches the size of the vessel being analysed. Therefore, the final estimate of vesselness is obtained by integrating the vesselness measure provided by the filter response at different scales,

$$V_0(\gamma) = \max_{s_{min} \leq s \leq s_{max}} V_0(s, \gamma) \tag{5.6}$$

where $s_{min}$ and $s_{max}$ are the minimum and maximum scales at which relevant structures are expected to be found, chosen so that they cover the range of relevant vessel widths.

Following the above enhancement procedure, a boost function is ap-

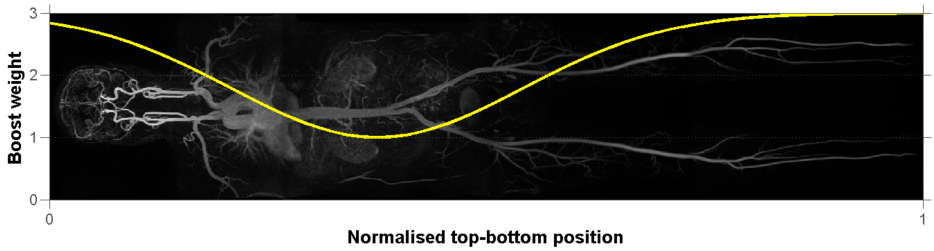



**Figure 5.7** – Visual representation of the "boost" function used for abdominal suppression and boosting signal in the extremities. (Source: [9])

plied [9]:

$$p_i = v_i \left( w - (w-1) \exp\left( -\frac{(x-c)^2}{2\sigma^2} \right) \right) \tag{5.7}$$

where $v_i$ is the vesselness value, $w$ controls the penalisation applied to the abdominal area and $c$ denotes the centre of the region, of size $\sigma$. This function was used as a method of suppressing residual signals left from organs in the abdominal cavity after subtraction and correcting for the loss of signal at the extremities which are typical due to difficulties in tracking the contrast bolus during the scan (Figure 5.7). The values were derived empirically by the authors and given as $w = 3$, $c = 0.35$, and $\sigma = \frac{1}{6}$ [9].

Finally, vascular tree is then segmented by applying a simple threshold to this enhanced dataset. Since the whole body scan consists of 4 separate stations, with possible intensity differences due to scan timing, the lower-leg scan is used to set the threshold value (lowest level of 4 stations), using $t = b/3$ where $b$ is the max vesselness value.

### Vessel Tracking via Centrelines

Arterial tracking in 3D is achieved through calculating the centrelines of the segmented vessels. Using the previously calculated binary segmentation, a skeletonisation methodology based on level sets and fast-marching techniques [4] is applied in order to obtain the corresponding centrelines with sub-pixel accuracy.

### Vessel Calibre Measurement

The GroBa system estimates the lumen calibre at each point along the centreline using a "balloon-growing" technique. Making the simplifying assumption that the fully inflated balloon has a cylindrical shape, this technique iteratively grows a balloon within the segmented vessel until

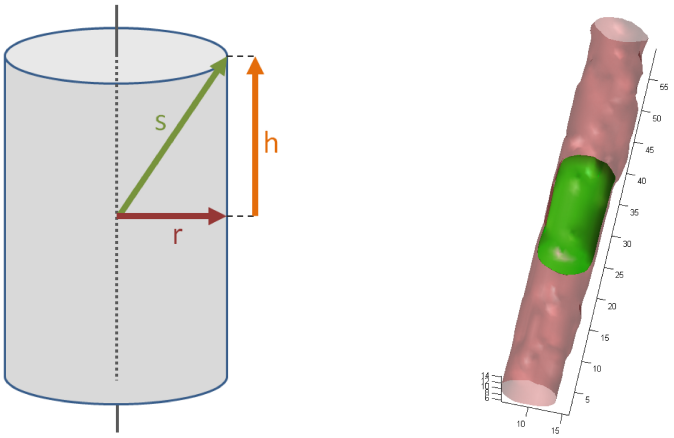



**Figure 5.8** – Visual representation of the inflated balloon within the segmented vessel (right). Using the cylindrical assumption the parameters $h$ and $r$ are unknown, while the half-diagonal $s$ and the volume $v$ of the balloon are known. (Source: [9])

the stopping criterion is reached; the height of the estimated balloon is equal to twice its calibre.

The process begins by initialising the balloon as a single voxel on the centreline, which is grown within the segmentation mask via binary dilation in 1-voxel steps. At each step the diameter and height of the balloon are estimated. First the position of the most distant voxel from the balloon centre (i.e. the initial centreline point) is found, and the euclidean distance between these two points is set as $s$, corresponding to the three-dimensional diagonal of the cylindrical balloon (as shown in Figure 5.8). The balloon diameter is then computed using the equation of a cylinder and the Pythagoras rule:

$$v = \pi r^2 2h \tag{5.8}$$

$$r^2 = s^2 - h^2 \tag{5.9}$$

where $v$ is the known volume (sum of all voxels), $r$ is the unknown cylinder radius, $h$ is the unknown half-height and $s$ is the known half-diagonal. From simple substitution and rearranging we get

$$v = 2\pi s^2 h - 2\pi h^3 \tag{5.10}$$

where now the only unknown is $h$, which can be obtained by finding the roots of Equation (5.11):

$$-2\pi h^3 + 2\pi s^2 h - v = 0 \tag{5.11}$$

The calibre of the vessel can then be estimated by assuming the balloon

at the final iteration approximates a cylinder, and therefore:

$$calibre = 2\sqrt{\frac{v}{2\pi h}} \tag{5.12}$$

The main effect of this approach is to essentially average over a small local region of the vessel, ensuring a more stable estimate of the calibre variation along the vessel in the presence of image noise or local artefacts. It should also enable more accurate calibre measurements where the estimated centrelines are noisy or otherwise inaccurate, and better cope with irregular, non-tubular vessel structures such as stenoses.

#### Measurement Visualisation

The final step in the presented pipeline is to display the calculated vessel calibres as an HSV colour overlay on the whole body MIP. This was chosen to theoretically allow easy visualisation of the vessel sizes throughout the body, with a sudden change in colour hue indicative of a local narrowing of the vessel. However, in our prototyping exercise carried out with clinicians (Section 3.3) it was found that this type of visualisation was not useful in a diagnostic setting, with point-markers being preferred, so we made no further examination or use of this visualisation approach.

### 5.3.2 GroBa Measurement Tests

Here we show the results of applying the GroBa system in a series of experiments with synthetic vessels to test its accuracy under ideal conditions. We also apply the system to a real WBMRA dataset, and compare a small number of sample ground truth stenoses against the GroBa results and a semi-automated measurement system called "VesselMetrix" (as described in Chapter 2).

We will first describe the data used.

#### Synthetic Lumina

Three types of synthetic vessel were created to test the mechanisms of the GroBa measurement system under controlled conditions. These binary vessel volumes consisted of the following morphologies:

**Synthetic Vessel 1** A simple cylinder with an idealised, symmetric stenosis on the central axis.

**Synthetic Vessel 2** A noisy vessel with a ∼33% stenosis and ∼20% aneurysm.

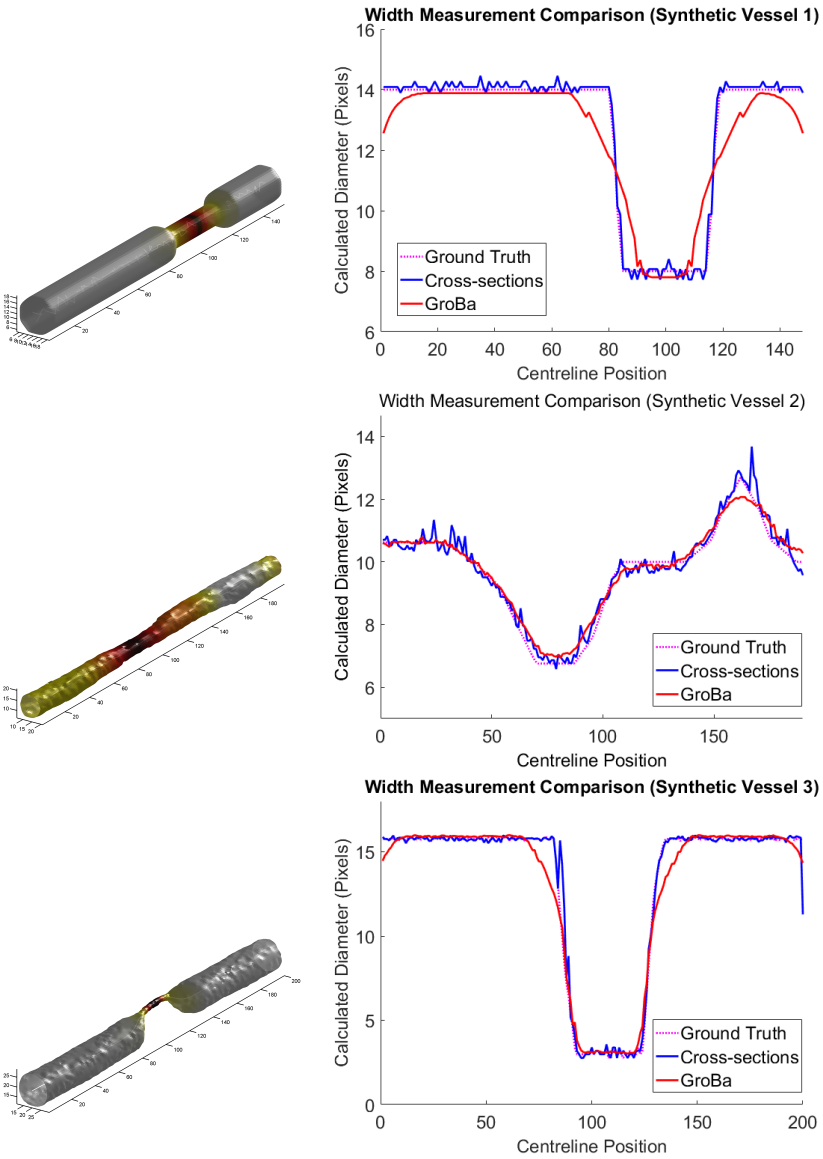



**Figure 5.9** – Illustration of the synthetic vessels (left), and corresponding graphs showing the calculated diameters using both the cross-section and GroBa technique.

**Synthetic Vessel 3** A wide vessel with a large, off-axis ∼80% stenosis.

Figure 5.9 shows a 3D rendering of the three vessel types, along with the corresponding graphs of the calculated vessel calibre. Also computed are the diameter estimates using the cross-section technique, whereby the area of a cross-sectional plane at each centreline point is calculated, oriented perpendicular to the vessel direction, and the equivalent vessel diameter computed by assuming a circular cross-section (i.e. $diameter = 2\sqrt{A/\pi}$ ).

We can see that for Case 1, the simplest vessel type, GroBa calculates the correct maximum and minimum diameters when compared to the known ground truth, providing a stable measure of the lumen calibre at the expense of having a smoother transition between the two calibres.

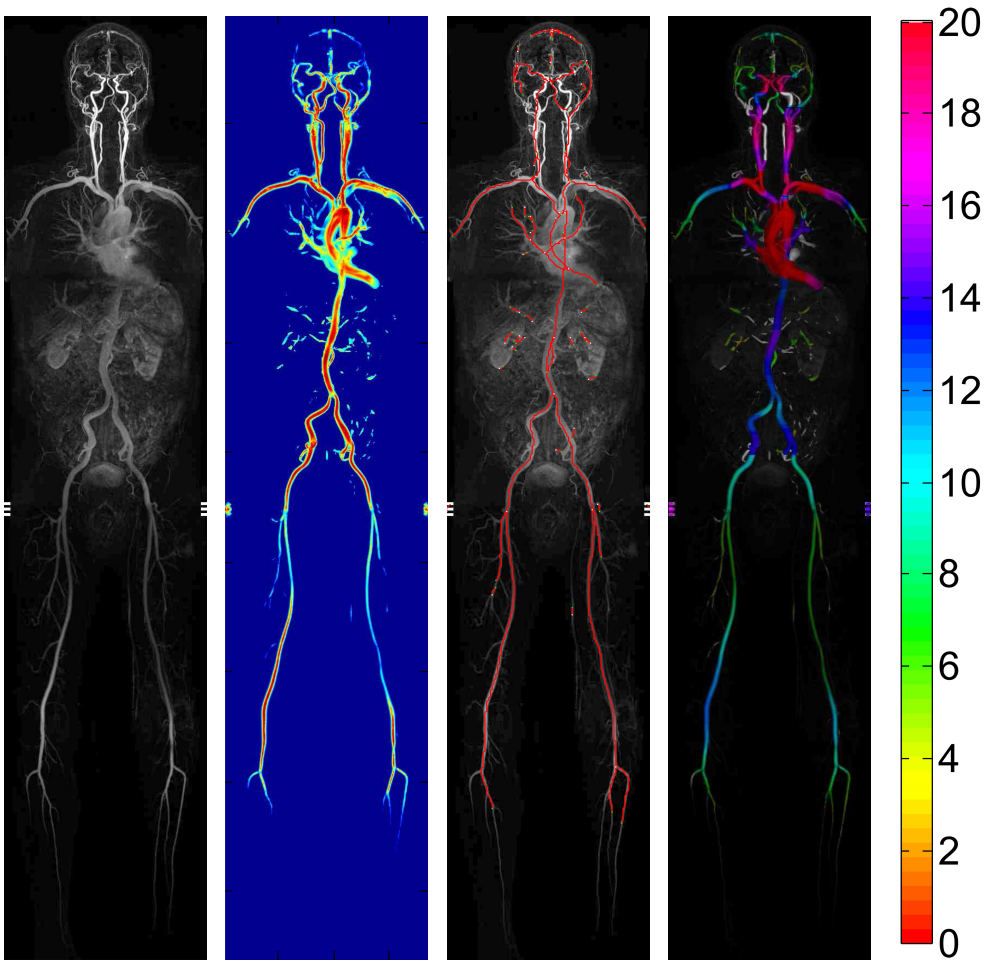



**Figure 5.10** – A WBMRA dataset processed using the GroBa system, showing (left to right) the original data, vessel enhancement (with abdominal suppression), calculated centrelines, and the HSV overlay representing the lumen calibre (with the scale measured in voxels).

Case 2 highlights how the technique agrees well with the ground truth, providing a more stable estimate as compared to the other two methods, but also shows that for rapid changes in diameter which occur over a spatially small section the smoothing effect also causes an inaccurate estimation of the true variation (underestimating the aneurysm by around 1 pixel in this case). Finally, Case 3 shows the technique to perform well for non-symmetric, off-axis stenoses.

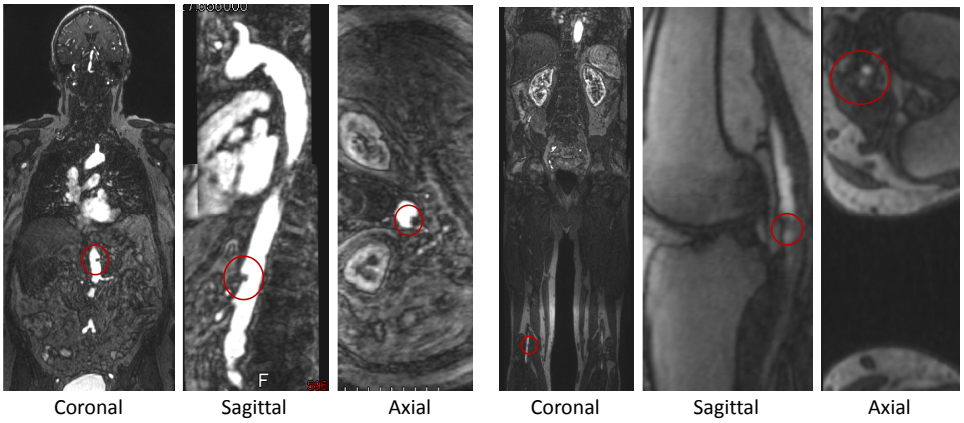
### Real WBMRA Datasets

The GroBa pipeline was applied to a test WBMRA scan, provided by Ninewells Hospital.

Figure 5.10 shows the outputs from each stage of the code. As can be seen, the code performs as described in Section 5.3.1; when presented with a new WBMRA dataset the vessels are first automatically enhanced using the Frangi vesselness filter, segmented via simple thresholding and

| Stenosis Grade | Equivalent Diameter Reduction |
|---|---|
| 0 | No Stenosis |
| 1 | 1–30% |
| 2 | 31–50% |
| 3 | 51–70% |
| 4 | 71–99% |
| 5 | 100% (Occlusion) |
| U | Non-diagnostic |

**Table 5.2** – Stenosis grade scale used in this project, as designed by the clinical team at Ninewells Hospital.




**(a)** Grade 1 **(b)** Grade 2

**Figure 5.11** – Annotated stenosis examples for grades 1 and 2. (a) shows the grade 1 (1-30%) stenosis in the abdominal aorta of the test patient. (b) shows the grade 2 (31-50%) stenosis in the right popliteal artery.

measured using the balloon-growing technique, with the final calculated lumen calibres displayed as an HSV colour overlay on the original dataset.

A brief evaluation was carried out to manually compare the calibre measurements obtained by the code on this real WBMRA dataset against stenosis grades assigned by clinicians. The WBMRA dataset provided for this evaluation was examined by a clinician at Ninewells Hospital, with stenoses assigned a grade between 1 and 5 based on the visible diameter reduction (see Table 5.2).

For this particular patient, two example stenoses were selected which were visibly clear, and could easily be measured with the VesselMetrix tool. These were assigned grades 1 and 2 on the severity score shown in Table 5.2.

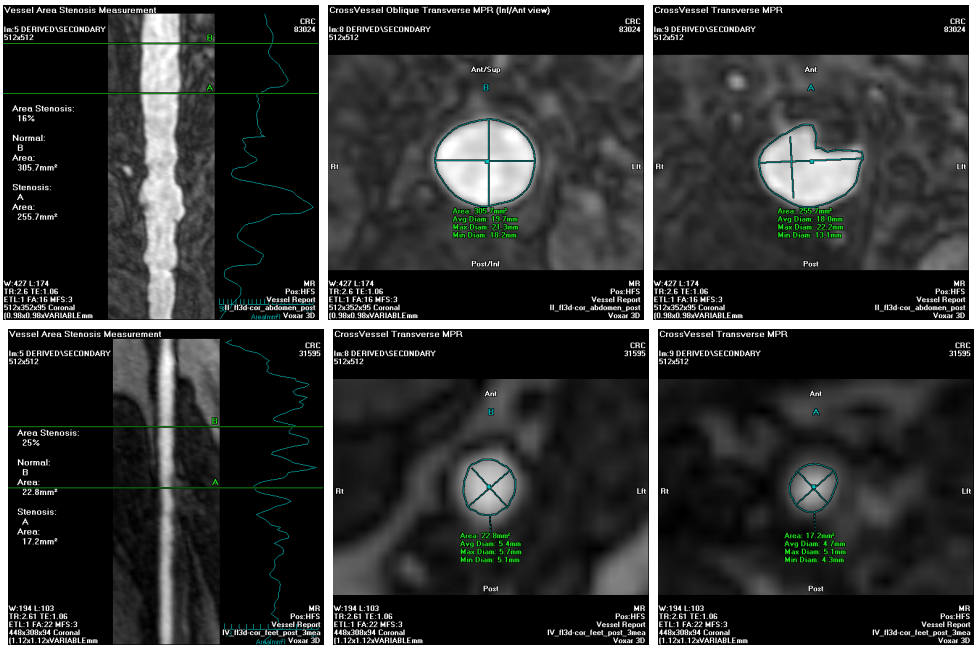



**Figure 5.12** – The top row shows the VesselMetrix measurements for the grade 1 stenosis, calculated as an area stenosis of 16%. The bottom row shows the measurements for the grade 2 stenosis, calculated to be an area stenosis of 25%.

The examples stenoses are shown in Figure 5.11. In each case the datasets were first opened in Voxar3D and the stenosis measured using the semi-automatic tools in VesselMetrix. The volume was then cropped and passed through the GroBa code. In order to handle the cropped volume correctly the abdominal suppression code was deactivated before processing. Finally, the centreline and corresponding widths were extracted, with the stenoses measured by averaging the "healthy" 5-8 measurements prior to the stenosis and dividing that by the minimum diameter found.

The results for the VesselMetrix measurements are shown in Figure 5.12, with the GroBa results shown in Figure 5.13.

It can be seen that for the grade 1 stenosis in the large abdominal aorta both VesselMetrix and GroBa correctly grade it to within the 1-30% range, measuring a 16% area and 23% calibre stenosis respectively. However, for the grade 2 stenosis, both methods appear to have difficulties, with VesselMetrix estimating the area stenosis as 25% and GroBa measuring the calibre stenosis as 15%, both placing them as grade 1.

### 5.3.3 Shortcomings of the Groba System

In this section we will examine the shortcomings of the GroBa system, exploring its sensitivity to various types of errors.

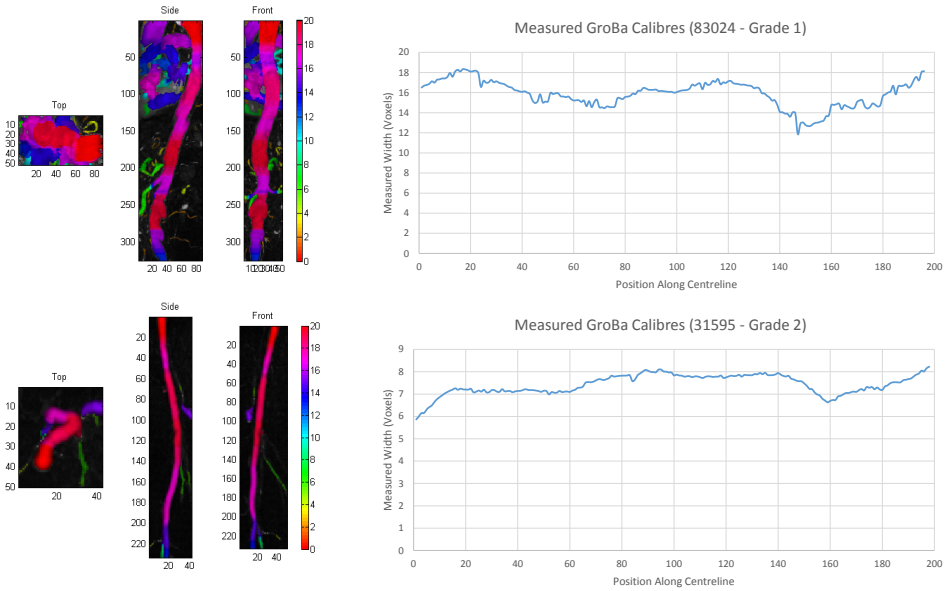



**Figure 5.13** – The top row shows the GroBa results for the grade 1 stenosis, with the bottom row showing the same results for the grade 2 stenosis. In each case the left image shows the measured calibres (in units of voxels) as an HSV colour overlay on the volume, with the graph showing the actual GroBa measurements along vessel. From these it was calculated that the grade 1 stenosis was 23% (at position 150) and the grade 2 stenosis was 15% (at position 160).

### Effect of the balloon's initialisation position on measured calibre

One of the key benefits of the balloon-growing technique for calibre measurement is that it should be stable to noisy centrelines, as opposed to the cross-section technique which is very sensitive to the estimation of vessel direction calculated from the centreline. The cross-section method is however stable to the position of the centreline within the vessel since it will return the same measurement regardless of where the centreline point is on the cross-sectional plane. It would therefore be desirous for the GroBa calibre measurement to return the same value regardless of where the balloon is initialised within the vessel.

To test GroBa's sensitivity to the seed position of the balloon we calculate the calibre of the synthetic vessels described in Section 5.3.2 by initialising the balloons at every voxel in a single cross-section (Figure 5.14a) and then every voxel in the entire volume (Figures 5.14b).

We can see from the Figure 5.14 that there is a broad area within the centre of the vessel which returns close to the same measurements regardless of the initialisation position. However, as the seed voxel moves closer to the vessel wall the measurements get smaller, with the worst case being on the vessel boundary itself. The reason for this appears to be due to the balloon not inflating fully as the stopping criterion (aspect

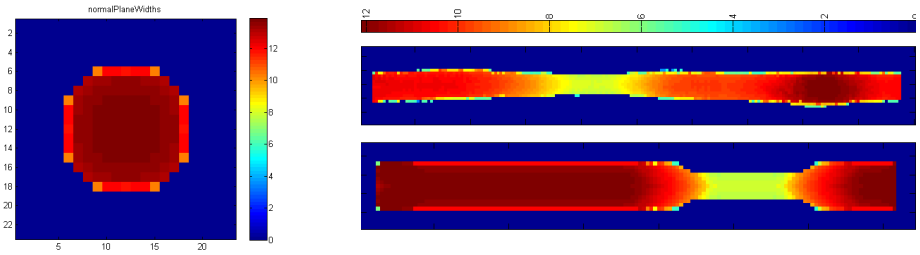



**Figure 5.14** – (a) shows the cross-section through a symmetric synthetic vessel (width 14 voxels), with the colour scale representing the calibre obtained by initialising the balloon at that voxel in the plane. (b) shows MIPs of two synthetic vessels shown in Section 5.3.2, with GroBa measurements initialised at every voxel in the vessel. Note that the calibre is consistently underestimated near the vessel boundary.

ratio 2:1) is triggered early due to non-symmetric expansion. Therefore the balloon does not inflate fully into the vessel, causing an underestimate of the true diameter.

Examples of balloons which have reached the stopping criterion when initialised at the central and edge voxels are shown in Figure 5.15.We can see that when the balloon is initialised very close to the vessel boundary the stopping criterion is reached before the balloon has fully inflated into the vessel. The same experiment was repeated using an increased height-to-width ratio of 3 (Figure 5.16). We can see that now the balloons inflate fully regardless of the initial seed voxel position. Furthermore, when applied to the entire vessel (shown in Figure 5.17) the edge-effect seen in Figure 5.14 is no longer present. Note however that increasing the height-to-width ratio increases the length of the balloon, and therefore the amount of vessel that is being averaged, thereby decreasing the method's overall accuracy and response to rapid changes in calibre. Also, the minimum ratio required to still allow the balloon to fully inflate will be dependant on the vessel calibre, so a true solution to this issue would likely involve using additional knowledge about the expected width of the vessel, or the formulation of a more complex set of stopping criteria.

### Effect of Vessel Junctions on Measured Calibre

It was recognised that the simple criterion used in the balloon expansion stage does no take into account any information about the structure of the vessel, and can therefore expand into branching vessels. Figure 5.18 shows the effect this has on the calculated calibre around vessel junctions. It can be seen that the calibre is overestimated on either side of the junction with an abrupt change seen downstream in main vessel, rep-

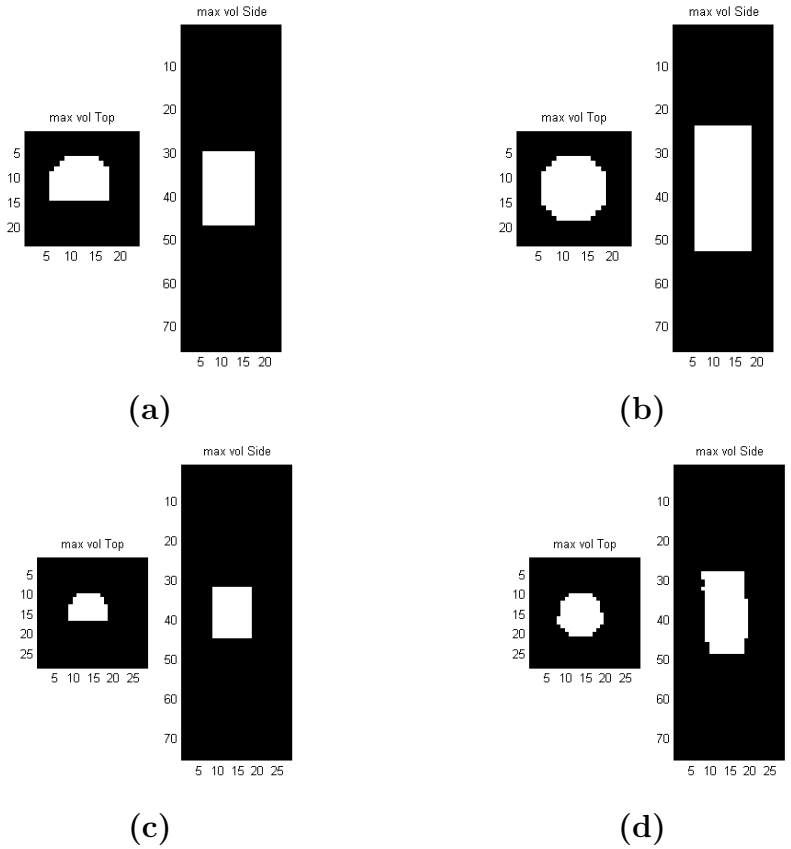


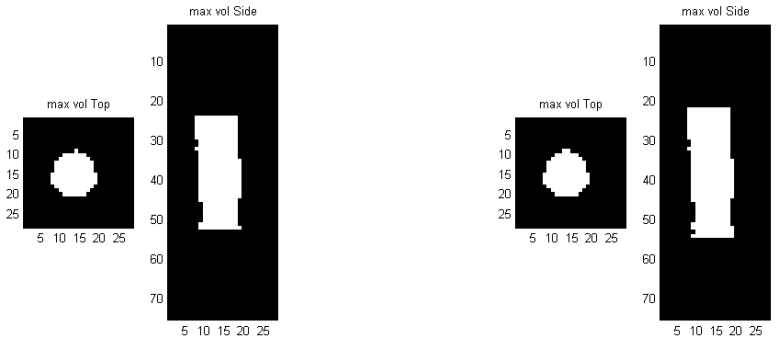

**Figure 5.15** – Axial and sagittal projections of the inflated balloons initialised at the edge (a and c) and centre (b and d) voxels, using the first (a and b) and second (c and d) synthetic lumina described in Section 5.3.2. In both cases the height-to-width ratio of the stopping criterion was 2.

resenting the point at which the stopping criterion is triggered before it can inflate into the wider junction and (potentially) the other branching vessel.

### 5.3.4 Effect of Vessel Direction with Respect To Inflation

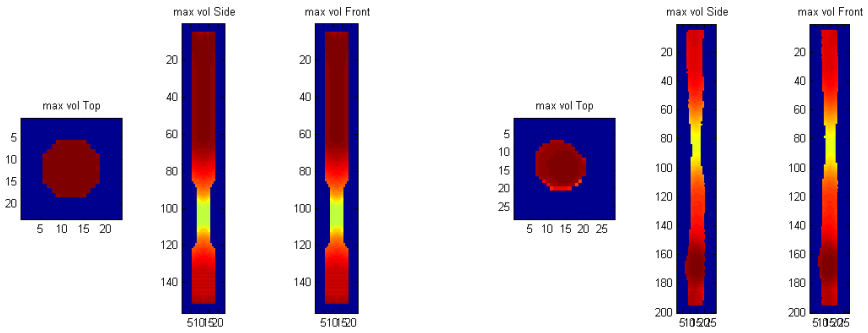
In the simple synthetic examples shown in Section 5.3.2, the vessel direction is roughly parallel to one of the x-,y-,z-axes of the dataset. It was noted that the dilation method used in the original GroBa implementation expanded symmetrically in the x-, y-, z-directions of the dataset, regardless of the vessel direction. This creates a growing cube from the centreline seed voxel, with balloon voxels which fall outside the segmented vessel being culled between each iteration.

To simulate the effect of a vessel running at 45 degrees to the direction of inflation, the balloon inflation itself was rotated 45 degrees and the previous experiment repeated for the first synthetic vessel type. Figure 5.19 shows the result. Comparing to the analogous results in Fig-




**(a)** **(b)**

**Figure 5.16** – Results from the same experiment described in Figure 5.15, showing axial and sagittal projections of the inflated balloon initialised at the edge (a) and centre (b) voxels with a height-to-width ratio of 3. Here, only the results for the second synthetic vessel are shown.




**(a)** **(b)**

**Figure 5.17** – MIPs of synthetic data, with GroBa measurements initialised at every voxel in the vessel with a height-to-width ratio stopping criterion of 3. Note that when compared to Figure 5.14, the calibre is no longer underestimated at the vessel boundary since the balloons inflate fully within the vessel, regardless of their seed position.

ures 5.14 & 5.15 we can see that the edge effect is much more pronounced (Figure 5.19c), and even when initialised at the centre of the vessel the calibre is underestimated by 1 voxel, or around 10%, due to the ends of the elongated balloon having an irregular shape and thereby breaking the cylindrical assumption which assumes the faces are flat.

### 5.3.5 GroBa Conclusions

The purpose of this study was to examine the "GroBa" method for lumen calibre measurement as a baseline for comparison to our own developed methods.

First, the code was confirmed to perform as designed using synthetic

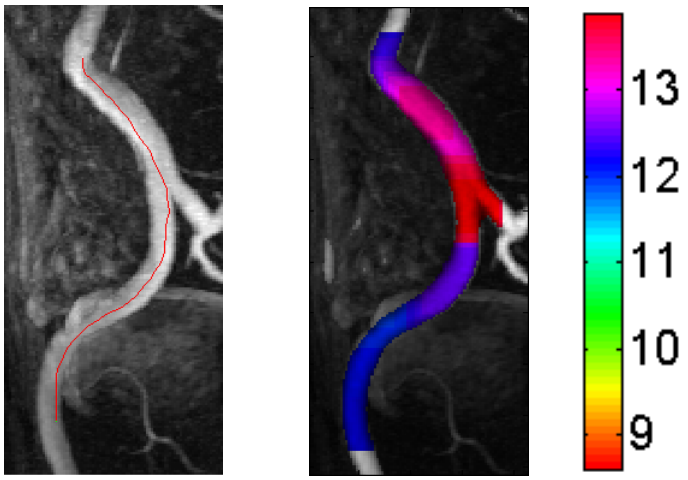
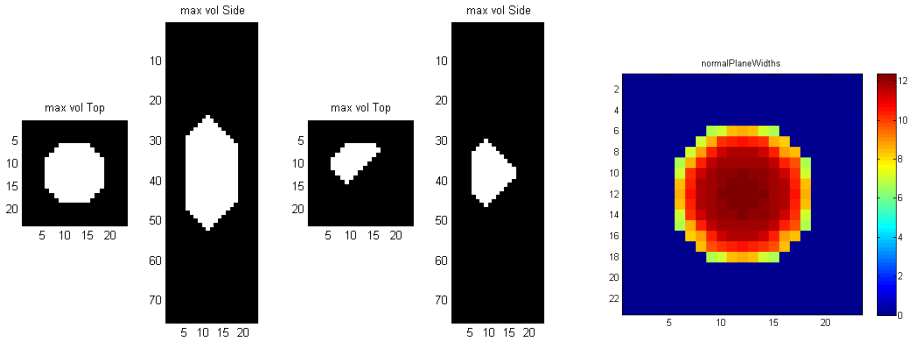



**Figure 5.18** – MIP of an artery extracted from the test patient dataset. The left image shows the computed centreline for the main vessel, while the right image shows the GroBa measurements (in voxels) computed along the centreline, overlaid in HSV color, demonstrating the effect of expanding into a junction off the main vessel.




**(a)** **(b)** **(c)**

**Figure 5.19** – Results of experiment where the synthetic vessel direction is at 45 degrees to the direction of expansion, showing the balloon shape when initialised on a central voxel (a), edge voxel (b), and finally a cross section of measured calibres initialised at every point in the plane (c).

and real WBMRA data (in the form of anonymised DICOM files). The calibre measurement technique was then examined in a series of synthetic tests to probe the reliability of the results. A positional bias was discovered, whereby if the balloon was initialised close to the vessel boundary the calibre would be underestimated. This was found to be mainly due to the asymmetric inflation of the balloon triggering the simple stopping criterion too early, before the balloon had fully inflated into the vessel.

Examining the measurements close to vessel junctions revealed that the calibre tended to be overestimated around these regions, though this is a shortcoming common to other techniques too.

The particular method used for inflating the balloon was also found to affect the results in vessels which didn't run parallel to the axes of the

dataset, causing the "faces" of the balloon to be irregularly shaped rather than flat, breaking the cylindrical assumption for the calibre estimation.

Finally, a comparison with ground truth annotations found that despite the discovered issues, the system was able to properly grade a small ($< 30\%$) stenosis in a large vessel. When the results of this automated system were compared with those achieved using the manual tools in VesselMetrix, it was found that both had similar performance (and encountered similar difficulties) in measuring the small number of stenoses examined here. A much broader comparison will be described in detail in Chapter 8.

Our conclusions are therefore that although the GroBa method does encounter some issues when applied directly to real stenosis examples, the pipeline that has been designed specifically for WBMRA vessel analysis provides a good framework to build upon, and the GroBa calibre method should be tested as a candidate measure for stenosis detection. To this end, we targeted improvements in two key areas highlighted by the prototyping exercise in Chapter 3.3.2; improvement to the vessel segmentation algorithm, validated on ground truth WBMRA segmentation maps (Chapter 6), and the development of stenosis grading algorithms, tested on ground truth stenosis annotations (Chapters 7 and 8).

# Chapter 6

# Automatic 3D Vessel Segmentation

## 6.1 Introduction

A critically important stage in our pipeline outlined in Figure 1.3 is the accurate segmentation of the arteries of interest. Segmentation of vascular structures is a common task to many medical applications [20, 97,98], and is a fundamental step in the the quantification of pathologies such as stenoses.

Many vessel segmentation techniques have been proposed in the literature, as explored in [20,21], and discussed in Chapter 2.3. In this chapter we examine three commonly used and well established techniques for vessel segmentation — active contours and two vesselness filters — comparing their results against our own method developed using a convolutional neural network (Convnet), structured as a voxel-wise binary classifier following the network structures used in other works such as [57]. This work was presented at MIUA 2017, and constituted the first publication of a deep learning segmentation network trained and tested on WBMRA data, and the first quantitative comparison of these vessel segmentation techniques for this application. Due to the poorer than expected performance of the 3D Convnet, we then extend the study to include a 2D U-Net architecture which we discuss in Section 6.6.

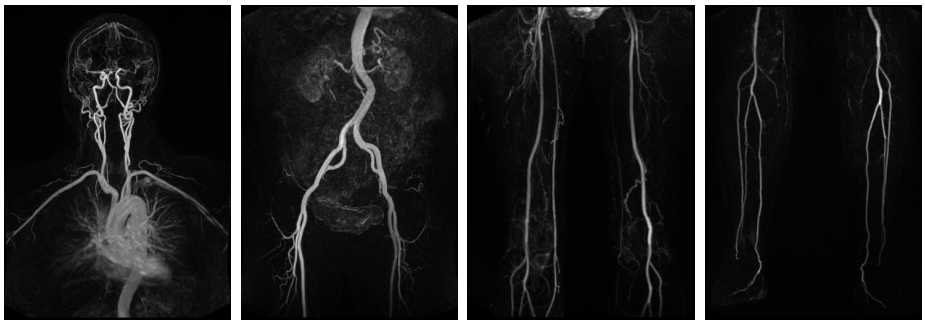


**Figure 6.1** – Maximum intensity projections of the four stations of patient 1, shown after digital subtraction of the pre-contrast from the post-contrast volumes.

## 6.2 Datasets and Protocols

### 6.2.1 Patient Data

The data selected for this study consists of three whole-body volumes, each of which are split into four "stations" as shown in Figure 6.1. These were acquired at Ninewells Hospital in Dundee, UK, using a 3.0 Tesla MRI scanner (Siemens Magnetom Trio), and are a subset of our full WBMRA dataset described in Section 4.2.

Due to time and resource constraints, the arteries in each individual station of 3 patient datasets were manually segmented by a single annotator. This was a Master-level student with a background in anatomy, who had previous experience annotating 2D X-ray images. Due to the shortcoming of having only a single annotator and being unable to perform any inter-rater variability analysis of the manual segmentations, the patient data was therefore chosen based on the visual quality of the scans after the pre-processing stages described in Section 4.5, with the three exhibiting the highest vessel contrast and lowest number of visible artefacts being chosen in order to simplify the extremely complex and time consuming segmentation task. To further aid with visualisation an intensity equalisation step was applied in the axial direction, ensuring a consistent vessel intensity across the entirety of each volume (detailed in Section 6.3.5).

### 6.2.2 Ground Truth Segmentation

The aim of collecting manual ground truth segmentations is to create a database of "true" vessel maps against which algorithms can be trained and tested, giving a fair assessment of their performance.

Boccardi, *et al.* [99] discuss the importance of a validated procedure

for segmenting anatomical features, in their case of the hippocampus. Examining the protocols for 12 different published studies, they found that there can be up to 2.5-fold difference in segmented volumes between different approaches. Another study by Joskowicz *et al.* [36, 100] found that segmentation volumes of a range of structures in CT (liver tumors, lung tumors, kidneys, and brain hematomas) had between a 25–50% variability in volume when examined by 11 radiologists following a strict protocol.

The main challenges of manual segmentation occur where contrast between the object of interest and the background is low, and where the object boundaries are smooth and not clearly defined. Therefore, it is imperative that a standardised protocol is followed during the data collection process, making clear what the aims of the task are and how decisions should be made regarding where the segmentation boundaries should be placed, in order to minimise the possible variation in the data not due to real variation in the object being annotated. No standardised protocol for the manual segmentation of arteries in MRA exists in the literature, so a bespoke protocol was developed as part of this study, as detailed below.

### 6.2.3 Segmentation Protocol: Step-by-Step

The open source software package "3DSlicer" [101] was selected for this task (version 4.5 at the time of use), as it was found to be easier to use by the annotator than some of the alternative programs such as ITK-SNAP and ImageJ, with the manual drawing tools providing an easy method of segmenting the vessels and the visualisation tools giving a clear view of the segmented surface in three dimensions.

The protocol was developed with guidance from the online training and user manual available from the 3D Slicer website, incorporating the use of the threshold tool as an effective starting point from which the manual delineations were created.

The final step-by-step protocol followed by the annotator for each data volume was as follows:

1. After pre-processing of the volumes as detailed in Section 6.2.1 the dataset was loaded into the 3D Slicer software.

2. The "Editor" module was used to create a new "label map volume", which stores the assigned label for each voxel in the volume

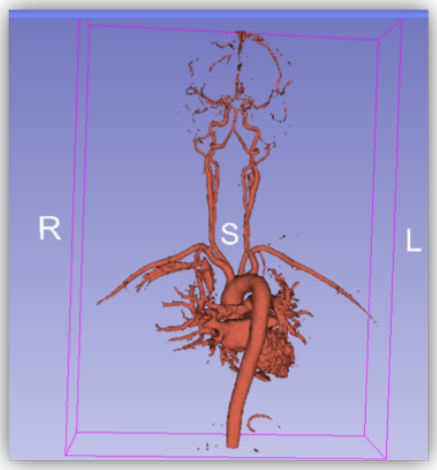



**Figure 6.2** – Example of isosurface created by MakeModelEffect for Station 1, giving a 3D rendering of the current segmentation volume.

3. Using the "Threshold Effect" tool, the software automatically detects a threshold. This was manually manipulated by the user according to each individual dataset. Viewed in cross-section, the software highlights the area corresponding to the current threshold as a colour overlay. This gives the user visual feedback on the effect of a change in threshold level, and was manipulated such that the contours across multiple slices aligned as closely as possible to those which would have been drawn manually. The criteria for this was that the contour should trace the mid-point of the greyscale transition between the bright vessel and dark background.

4. To visualise the current segmentation, the "MakeModelEffect" tool was used. This creates a 3D representation of the label volume and allows the user to quickly identify any mislabelled regions. An example rendering of one such case is shown in Figure 6.2. It is important to observe that smaller vessels do not appear complete and must be manually manipulated. In addition to this, details of the heart, lungs and associated vessels have been captured and must be manually corrected.

5. To manually label any regions of artery which have been incorrectly labelled as background, the "PaintEffect" tool was used. This acts like a paintbrush in standard image editing programs, and allows the user to manually label voxels as belonging to the current label volume. So the user can add sections of vessel which were missed by the threshold or otherwise inaccurate or incomplete.

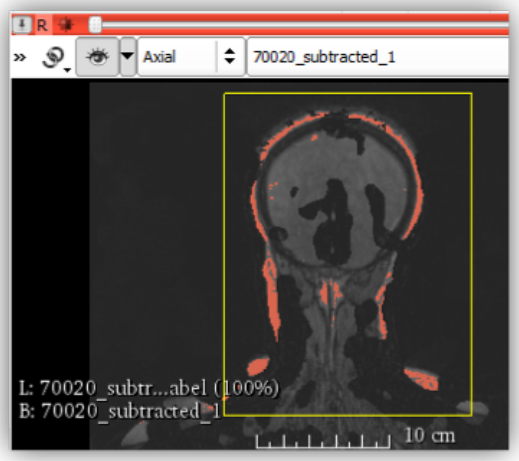



**Figure 6.3** – Example incorrectly labeled voxels around the outside of the skull being removed using the "RectangleEffect" option of the "EraseLabel" tool.

6. To correct any regions that were labelled as artery that are background voxels or belong to an irrelevant anatomical structure, the "EraseLabel" tool was used. This can operate identically to the "PaintEffect" tool, assigning the clicked voxels to the background label, or using a rectangular selection tool to delete all enclosed voxels at once from the segmentation (Figure 6.3).

7. Once the initial model has been generated, each slice was manually checked by eye to ensure the entire artery has been labelled correctly (i.e. not over-segmented or under-segmented). Any modifications were made in the slice views (not 3D view).

8. Finally, the segmentation is considered complete if there is a continuous path through each of the segmented arteries. In addition, it was determined that once arteries reached less than 2 voxels in diameter or if sections of the arteries cannot be clearly identified, that they are discounted from the segmentation as being too unreliable to label. An example of a completed segmentation volume is shown in Figure 6.4.

Following the above procedure three patient datasets were manually segmented over the course of eight weeks. Coronal MIPs of all resulting segmentation maps are shown in Figure 6.5.

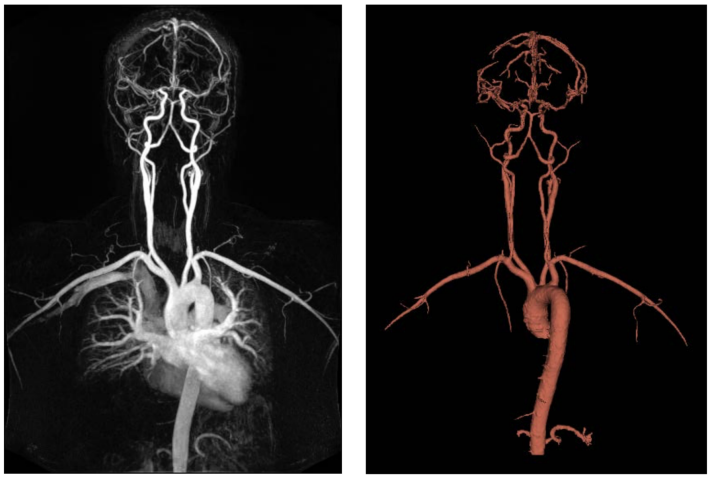


**Figure 6.4** – MIP of the first station of a patient dataset (left) and the corresponding artery segmentation (right) after manual labelling.

## 6.3 Segmentation Methods

### 6.3.1 Active Contours

Segmentation using active contours, where an initial curve is evolved using a cost function depending on local gradients (external forces) and shape constraints (internal forces), was first proposed in [102] and has been successfully applied to many segmentation problems [21, 103]. In general, active contour methods are based on image gradient, detecting edges and generating well-defined boundaries on which to evaluate the internal and external energy.

The level set model tries to solve an optimisation problem defined by embedding the active contour as a constant set (zero level) in a surface $\phi$ that evolves in time with speed $S$, guided by the image gradients.

For our comparative study we chose the classic Chan-Vese model [103]. It has been applied to the segmentation of objects whose edges are not well defined by the gradient, and has a well defined implementation for 3D segmentation as described in [104].

The Chan-Vese model is formulated as a "mean-curvature flow"-like evolving active contour, where the stopping term depends not on the gradient of the image, as in classical active contour models, but is instead related to a particular segmentation of the image [103].

For 3D data, we define the bounded domain $\Omega \in \mathbb{R}^n$ (in our case $n = 3$), and the bounded image function $I : \Omega \to \mathbb{R}$. $\Omega$ can be divided into a set of connected domains by a curve $C$ by $\Omega - C = \cup_{i \in I} \Omega_i$. We then define two different regions $R_1 = \cup_{i \in I_1} \Omega_i$ and $R_2 = \cup_{i \in I_2} \Omega_i$ that represent the object support and the background support respectively.

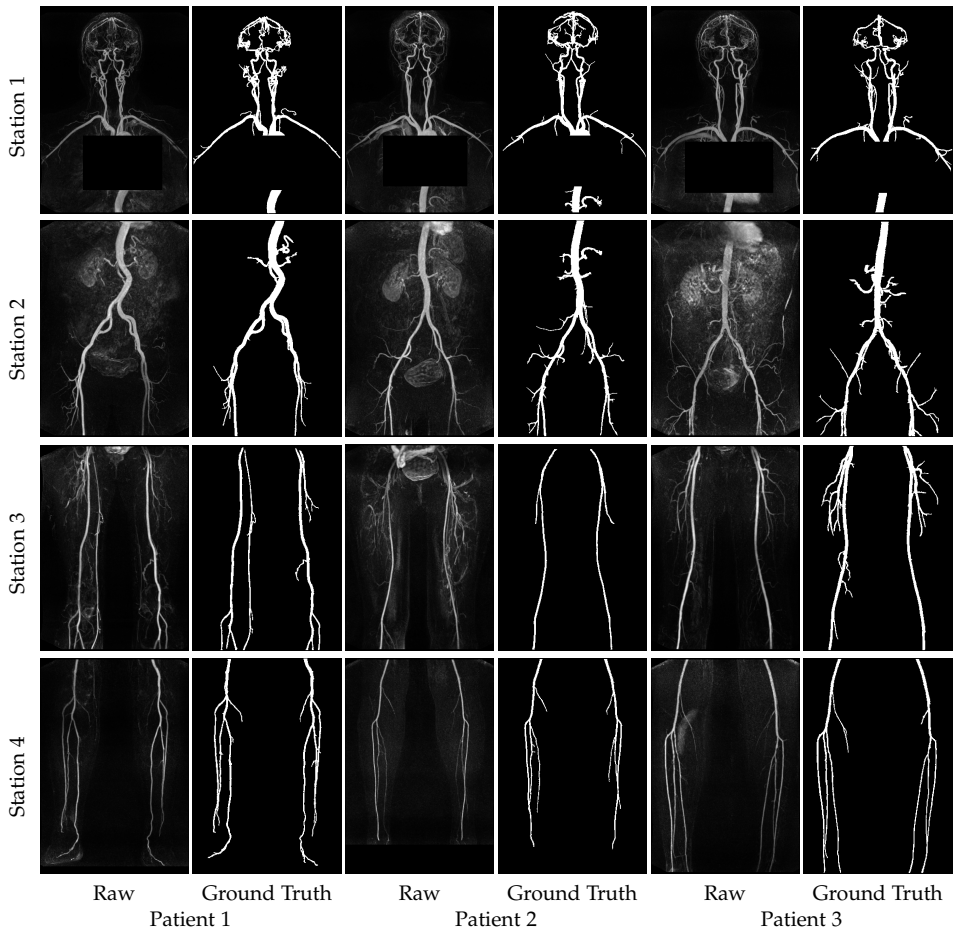



**Figure 6.5** – Coronal MIPs of the three raw patient datasets and their corresponding ground truth vessel maps, acquired using the described annotation protocol.

The final energy functional $E(\phi, \mu_1, \mu_2)$ is then given by

$$\begin{aligned} E(\phi, \mu_1, \mu_2) = \quad & \lambda_1 \int_\Omega (I - \mu_1)^2 H(\phi) d\Omega \\ & + \lambda_2 \int_\Omega (I - \mu_2)^2 (1 - H(\phi)) d\Omega \\ & + \alpha \int_\Omega H(\phi) d\Omega \\ & + \beta \int_\Omega | \bigtriangledown H(\phi)| d\Omega \end{aligned} \tag{6.1}$$

where $\mu_1$ and $\mu_2$ represent the mean value of the object support region and background support region of image $I$ respectively, and $H(\phi)$ is the Heaviside function. Here, the first two terms measure the variations around the mean intensity inside and outside the active contour, the third term measures the area inside the contour and the fourth term measures the length of the contour [104].

### 6.3.2 Vessel Enhancement Filters

#### Frangi Filter

This classic method of enhancing vessel-like structures is based on calculating the local curvature by analysing the Hessian function [44], and has been used extensively for vessel segmentation tasks across many imaging modalities, including the GroBa method for WBMRA data. It was therefore chosen as the baseline filtering method for this comparison study, and a full description of the method can be found in Section 5.3.1.

#### Optimally Oriented Flux

The "optimally oriented flux" filter, first published in [48], evaluates a scalar measure of the flux flowing through a spherical surface. Before computing this value, directional information is extracted by projecting the gradient along "optimal" axes, and the flux measure then evaluated. For each voxel a sphere with variable radius is built, centred on the voxel, which produces an "OOF response" when touching an object edge. If the voxel is inside the curvilinear structure the response will be positive, otherwise it will be negative.

The outwardly oriented flux along the direction $\hat{\rho}$ is firstly computed by projecting the gradient $\mathbf{v}$ along $\hat{\rho}$, with the flux then evaluated through the spherical region $S_r$ with radius $r$ using the definition

$$f(\mathbf{x}; \mathbf{r}, \hat{\rho}) = \int_{\delta S r} ((\mathbf{v}(\mathbf{x} + \mathbf{h}) \cdot \hat{\rho})\hat{\rho}) \cdot \hat{n}\, dA \tag{6.2}$$

where $dA$ is the infinitesimal area of $S_r$, $\hat{n}$ is the unit normal to the surface at position $h = r\hat{n}$.

As before, the goal is to obtain the principal eigenvalues of the intensities in a neighbourhood of each voxel. Inside the vessel, when the local spherical region with surface $S_r$ touches the boundaries of the object, $\mathbf{v}$ is oriented opposite to the direction of $\hat{n}$, therefore the eigenvalues $\lambda_1 \leq \lambda_2 \ll 0$. The gradient of the image will be perpendicular to the direction of the curvilinear structure, with a value of $\lambda_3 \approx 0$. In the case where the voxel is in the background, $\mathbf{v}$ will have the same direction as $\hat{n}$, and therefore $\lambda_3 \gg 0$.

To obtain the maximum response to the OOF while changing the radius $r$, we evaluate the geometric mean of the eigenvalues, as

$$M(\mathbf{x};s) = \begin{cases} \sqrt{|\lambda_1(\mathbf{x},s)\lambda_2(\mathbf{x},s)|} & \lambda_1(\mathbf{x},s) \leq \lambda_2(\mathbf{x},s) < 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.3}$$

where $s$ represents the scale factor. Similar to the Frangi approach, selecting the maximum response over an appropriate range of scales generates the final map, which can be thresholded to produce a final segmentation.

### 6.3.3 Convolutional Neural Network

In recent years, deep Convnet approaches have been driving advances in many computer vision tasks, such as image classification [56, 57] and image segmentation [58, 59]. Many network models have been developed for these tasks, and it is a very active area of research [60]. The network structure we chose was inspired by those explored in [57], and recently applied to segmentation tasks in MRI [105–107]. To the best of our knowledge, this is the reported results of applying a Convnet to vessel segmentation in WBMRA, for which no public sets of manually annotated vascular networks currently exist.

The final network structure is shown in Fig. 6.6, consisting of five layers; 3 sets of convolutional and max-pooling layers, followed by two fully connected layers. The output node of the final layer gives a single binary output of vessel/non-vessel for the central voxel of the input patch. This structure was arrived at after testing many deeper and shallower networks, and it was found that a 5-layer network gave the best performance given the small amount of data we had available for training.

Our network was implemented using Keras v1.1.0 and Theano v0.8.2. All layer activation functions were "ReLU" except the final output node, which was "sigmoid". The "Adam" optimiser was used during training, with "binary cross-entropy" selected as the loss function [108].

While this approach to voxel-wise classification has been shown to be less computationally efficient than a fully convolutional network [59], it allowed fine control over dataset balancing for our limited amount of ground truth data, and the fully connected layer gave additional flexibility to the network without increasing the required input volume size, which is inherent to the operation of convolutional layers.

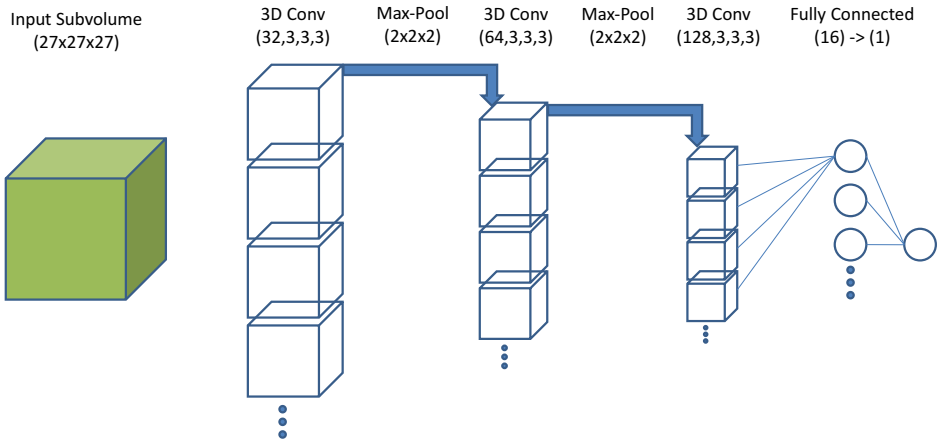



**Figure 6.6** – Structure of the 3D Convnet segmentation network. All layer activation functions were ReLU except the final output node, which was sigmoid. The Adam optimiser was used during training, with binary cross-entropy selected as the loss function

### 6.3.4 Comparison Criterion: Dice Coefficient

There are many metrics used for evaluating the quality of segmentation in medical images [109]. For our data, we have selected the Dice Similarity Coefficient (DSC — also referred to as the F1-Measure). This is given by

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} \tag{6.4}$$

where $|X|$ is the number of all the vessel voxels in the segmentation obtained by the tested method and $|Y|$ is the number of all the vessel voxels in the ground truth.

### 6.3.5 Pre-Processing and Parameter Optimisation

#### Pre-Processing

As can be seen in Figure 4.1, the subtraction of the pre-contrast from the post-contrast volume still leaves some tissues and non-arterial structures behind, particularly in stations 1 and 2. The most problematic of these are the lungs in station 1, which contain vessels which were not included in the manual ground truth. For this reason, the small region around the lungs and heart were masked out in both the original volumes and the ground truth data, excluding this area from our analyses.

Another artefact which remained after subtraction was the variation of lumen intensity along the length of vessels. These may arise due to poor timing of the contrast agent during acquisition, or by inhomogeneities in the magnetic field (such as surface coil artefacts). A simple

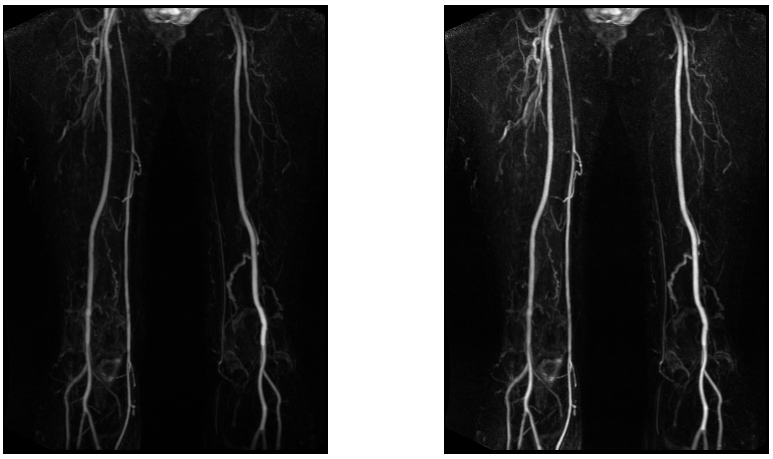


**Figure 6.7** – Results of intensity equalisation on station 3 of patient 1. The MIP of the raw volume is show on the left, and the equalised volume on the right.

procedure was followed to correct these variations in each station, in the form of a local intensity normalisation.

First we make the assumption that each axial slice contains an artery, and they are the brightest objects present (which holds true for all regions except for slices above the head and below the feet — these slices were simple masked to zero after the procedure was applied). We then applied a 7-slice sliding window axially, in which the local vessel intensity was estimated from its histogram by choosing the highest frequency bin above 70% of the maximum intensity, with this value corresponding to the vessel intensity estimate for the central slice. Once calculated for the entire volume, Gaussian smoothing of the values was applied and then each slice divided by its corresponding estimate. An example of the results of this processing is show in Figure 6.7.

### Active Contours

For the active contour method, the Toolbox implementation provided by [104] was used. Values for the the smoothing weight term, image weight term, and time step were fixed using a grid search optimisation procedure across all patients, with $\beta = 0.08$, $\Delta t = 2.72$, $\lambda = 0.0002$.

The final step was the initialisation of $\phi_0$. This choice was critical as it affects the time and the speed of the evolution of the curve. So again under the hypothesis that the highest intensity voxels belong to the vessels, we took a set of seed points with high grey levels as $\phi_0$. To keep the process completely automatic we used the multi-threshold formulation of Otsu's method for generating thresholds from grey-level histograms [110]. This was done using the Matlab "multithresh" function, which generates

| Station Number | SF / Radii | $\alpha$ (Frangi) | $\beta$ (Frangi) | $\sigma$ (OOF) |
|---|---|---|---|---|
| 1 | 1:1:15 | 0.5 | 0.5 | 0.4 |
| 2 | 1:1:10 | 0.5 | 0.5 | 0.5 |
| 3 | 3:0.25:5 | 0.5 | 0.5 | 0.5 |
| 4 | 1:0.5:4 | 0.5 | 0.5 | 0.4 |

**Table 6.1** – Enhancement filter parameters. The scale factor (SF) and radii values are written in the form *minimum:step:maximum*.

$N-1$ thresholds that maximise the inter-class variance between $N$ different classes (to a maximum of 20 for the Matlab function).

For each station we generated 10 thresholds on a patient pair, which served as 10 different sets of seed points. The active contour method was then applied using the above parameters, and the threshold generating the highest Dice score recorded. This threshold was then applied to the held-out volume to generate the initialisation of $\phi_0$, and the corresponding Dice score recorded for that patient and station.

### Enhancement Filters

The optimal parameters for the enhancement filters are shown in Table 6.1. These were optimised for each station across all patients using a grid search, with a fixed segmentation threshold.

The final segmentations were acquired by calculating the vesselness map using the parameters in Table 6.1, then 20 thresholds automatically calculated on each patient pair using the "multithresh" function described previously. The threshold achieving the highest Dice score was then applied to the held-out volume and the Dice score recorded.

### Network Meta-Parameters

A number of network structures were explored during optimisation of the network structure. Inspired by models discussed in [57] and [59], we trained models consisting of 2 – 6 convolutional layers with 16 – 128 $3 \times 3 \times 3$ kernels, 1 – 3 max-pooling layers, and 1 – 2 fully-connected layers. To help combat overfitting, $l_2$ weight regularisation was used for each convolutional layer [108], and 20% dropout used on the fully connected layers [111]. All layer weights were initialised from a scaled Gaussian distribution.

A single network was trained for each station, with training patches

extracted from two patients and the trained model applied to the held-out third patient in a 3-fold cross-validation setup.

The cubic patches were varied in size according to the network structure used, based on ensuring that the deepest layer still received a patch large enough to perform meaningful calculations on. A minimum side length of 15 voxels was needed to capture the thickest vessels, leading to a side length range of 15 – 50 voxels for the network structures we explored. For our final network, a patch size of $27 \times 27 \times 27$ was found to be optimal.

Finally, the number of training patches was chosen to maximise the available data. For each station, the minimum number of ground truth vessel voxels across all 3 patients was calculated, and this used as the number of positive samples to be extracted from each patient.

The data was balanced by extracting an equal amount of background samples. The position of the background samples were weighted to have two-thirds from regions within 5 voxels of a vessel and one third sampled randomly from the rest of the volume. This was found to improve the networks tendency to over estimate the diameter of the vessels when the background patches were sampled completely at random.

Data augmentation tests were carried out using rotation of the volumes around the three primary axes, and horizontal mirroring (swapping left and right sides of the body). Only rotation along the axial plane (i.e. rotating around the vertical axis) was found to improve performance for our network and data.

The total number of training samples used for each station were 136000 for station 1, 160000 for station 2, 28000 for station 3, and finally 24000 for station 4. During training, 5% of the training was data held out for validation, and the best network weights saved as those giving the highest validation accuracy score after 20 epochs (the network performance was found to typically converge after 8 – 12 epochs).

The networks were trained using an Nvidia Titan X Pascal GPU, with training times of between 3 – 4 hours for each model (depending on the station and number of training samples used).

## 6.4 Results

The results of applying the four automated strategies described in Section 6.3 to all stations and patients are shown in Table 6.3, with the corresponding segmentation results of patient 1 shown in Figure 6.8.

| Augmentation | Station | DSC | Mean DSC |
|---|---|---|---|
| None | 1 | 0.727 | 0.557 |
| | 2 | 0.543 | |
| | 3 | 0.231 | |
| | 4 | 0.727 | |
| Mirroring | 1 | 0.702 | 0.490 |
| | 2 | 0.550 | |
| | 3 | 0.232 | |
| | 4 | 0.477 | |
| Rotation (Axial) | 1 | 0.752 | **0.575** |
| | 2 | 0.552 | |
| | 3 | 0.303 | |
| | 4 | 0.692 | |
| Rotation (Sagittal) | 1 | 0.743 | 0.515 |
| | 2 | 0.518 | |
| | 3 | 0.275 | |
| | 4 | 0.522 | |
| Rotation (Coronal) | 1 | 0.776 | 0.465 |
| | 2 | 0.487 | |
| | 3 | 0.260 | |
| | 4 | 0.339 | |
| All Rotations & Mirroring | 1 | 0.730 | 0.528 |
| | 2 | 0.522 | |
| | 3 | 0.288 | |
| | 4 | 0.572 | |

**Table 6.2** – Dice coefficients for after augmentation for the worst performing dataset (patient 2).

It can be seen from Table 6.3 that for our case of only three patients, the OOF filter achieved the greatest mean DSC of 0.705. The 3D Convnet typically outperforms at least one of the other techniques, except for station 2. The main reason for this appears to be because of the additional artefacts left over from the imperfect volume registration and subtraction procedure (particularly the kidneys and bladder). The network approach had the most difficulty distinguishing between these

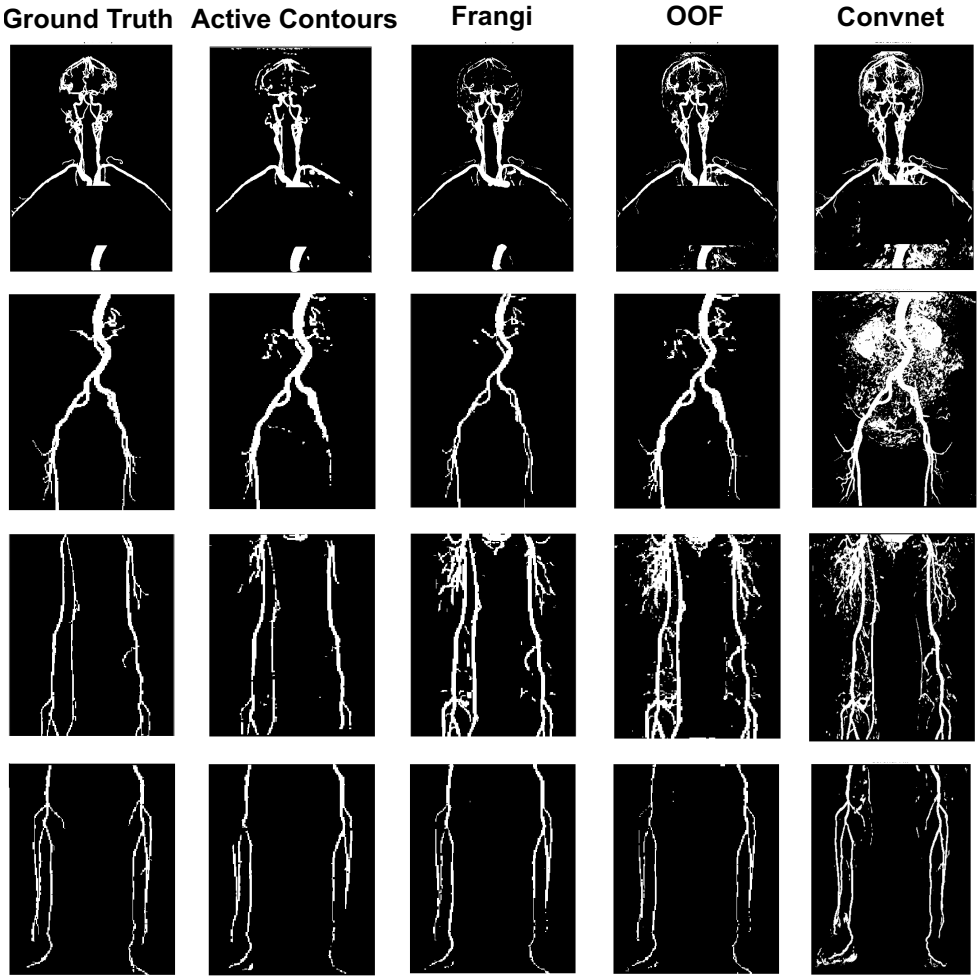



**Figure 6.8** – Segmentation results for patient 1, shown as coronal projections.

artefacts and the arteries, causing it to over-segment station 2, resulting in a lower Dice score.

Looking at the segmentation results in Figure 6.8, a number of observations can be made. The active contour method often produces broken vessels, such as the right branch in station 2, and has the most difficulty segmenting the finest vessels in station 4.

The Frangi and OOF enhancement filters produce visually similar results, though the OOF performs better at rejecting non-vessel artefacts (most noticeably in the brain and abdomen of station 1). Both filters do exhibit difficulties segmenting the finest vessels, such as at the bottom of station 2, and in different cases tend to either underestimate (Frangi in station 2, OOF in station 4) or overestimate (Frangi and OOF in station 3) the true diameter of the vessels as compared to the ground truth.

The Convnet performs poorest at rejecting non-arterial artefacts in station 2. This is likely due to it being a relatively shallow network, which we recognise as a shortcoming of the particular network architec-

| Patient (Station) | Level Set | Frangi | OOF | 3D Convnet |
|---|---|---|---|---|
| 1 (1) | 0.785 | 0.697 | **0.795** | 0.727 |
| 1 (2) | **0.837** | 0.817 | 0.812 | 0.618 |
| 1 (3) | 0.647 | **0.702** | 0.657 | 0.605 |
| 1 (4) | 0.672 | 0.780 | **0.843** | 0.806 |
| 2 (1) | **0.803** | 0.689 | 0.794 | 0.752 |
| 2 (2) | 0.712 | **0.722** | 0.674 | 0.552 |
| 2 (3) | **0.353** | 0.323 | 0.281 | 0.303 |
| 2 (4) | 0.662 | 0.776 | **0.844** | 0.692 |
| 3 (1) | 0.572 | **0.665** | 0.663 | 0.567 |
| 3 (2) | **0.504** | 0.488 | 0.469 | 0.422 |
| 3 (3) | 0.803 | 0.747 | 0.835 | **0.845** |
| 3 (4) | 0.618 | **0.843** | 0.791 | 0.697 |
| Mean DSC | 0.660 | 0.690 | **0.705** | 0.632 |

**Table 6.3** – Dice Coefficients for each method

ture tested, with the lack of training data preventing us from training a deeper network from scratch. It did however appear to have the highest sensitivity to extracting fine, low contrast vessels as can be seen in the lower half of stations 2 and 3.

## 6.5 Conclusions

In the above study we presented a quantitative comparison between four automated vessel segmentation technique for WBMRA data, using three manually segmented patient datasets. In this regime of limited ground truth data, it has been found that the Optimally Oriented Flux filter performed the best, with a mean DSC of 0.705. Visually, the Convnet approach segments vessels most consistently, with the least number of breaks, picking up finer vessels, and having the most consistently accurate diameters when compared with the ground truth. However it performed poorest at rejecting non-arterial artefacts, resulting in a lower DSC overall. It was also noted that some of the fine vessels segmented by the Convnet were not present in our ground truth. Due to having ground truth from a single observer, we are unable to estimate the quality and reliability of the ground truth data, and therefore the impact of

this on the DSC results cannot be easily estimated for our data. We are not aware of any publicly available sets of manually annotated vascular networks for WBMRA volumes.

The Convnet approach appears to be mainly limited by the lack of training data, which prevented a deeper network from being properly trained. One approach often used to ameliorate this issue is to fine-tune a previously trained network such as GoogLeNet, a 22 layer network trained on a database of 1 million natural images [112]. However, currently there are no pre-trained 3D networks available for medical data, and thus the training of deeper networks will require either a much larger database of ground truth segmentations, or a different network architecture which can be more easily trained with fewer annotations such as a U-Net [58].

## 6.6 Vessel Segmentation with a Fully Convolutional Neural Network

### 6.6.1 Introduction

The above comparison study was published at MIUA 2017 [113], and the OOF method was used to produce all 18 segmentation maps used for the analyses in Chapter 8. However, due to the poor performance of the Convnet proposed in Section 6.3.3 we subsequently extended this comparison to an additional network structure, namely the U-Net architecture first published by Ronneberger *et al.* in [58].

In this section we will describe the network and parameters used, compare the results against the previous methods, and discuss other possible extensions to this work.

### 6.6.2 The U-Net Architecture

The U-Net structure was first proposed in [58] for use in biomedical image segmentation. It is a "fully convolutional network" (using only convolutional layers), and employs a symmetric encoder-decoder like structure whereby a contracting path is used to capture context and a symmetric expanding path enables precise localisation.

The architecture was designed with particular regards to improving performance with very few training images, with the usual contracting network supplemented by successive layers with upsampling operators,

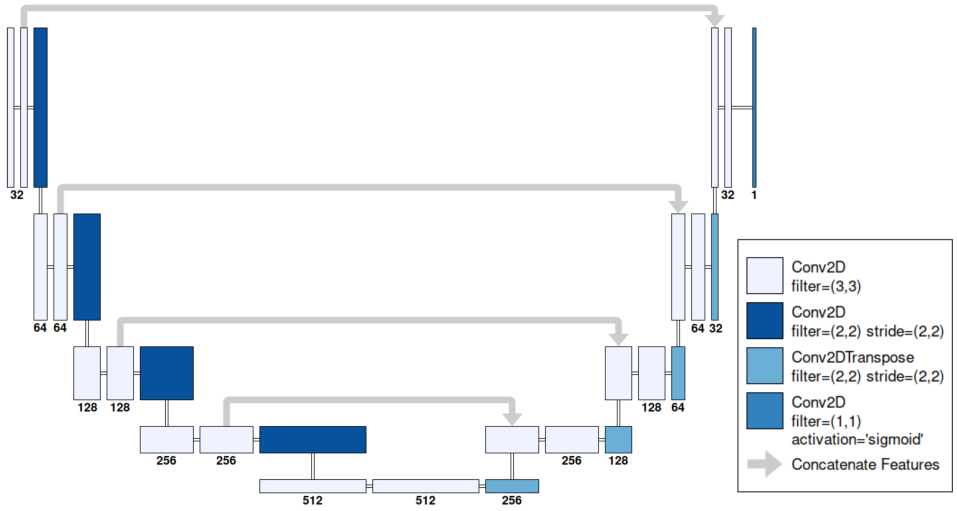



**Figure 6.9** – The U-Net network architecture designed for our segmentation task. All layers use same padding, ReLU activation function and a stride of (1,1) unless otherwise stated. The number under each layer indicates the number of features used at each step.

increasing the resolution of the output. Localisation performance was further increased by combining high resolution features from the contracting path with the upsampled layers at the same level.

A schematic of our U-Net network structure is shown in Figure 6.9. The network consists of a series of 2D convolutional layers, with all layers using the ReLU activation function except for the final output which uses a sigmoid. Each convolution layer supports inputs with up to three channels for handling RGB images. Since our data contains only a single channel for intensity but is 3D in nature, we tested the network with two different inputs; firstly using a single input channel, and secondly using three slices. This allows us to compare the performance of the exact same network structure when operating on a single 2D slice and three slices at once, introducing some additional 3D spatial context.

### 6.6.3 Training

The training process followed the same strategy as for the 3D Convnet, being split on a station-by-station basis using two patient volumes for training and validation, before testing on the held out patient. Training samples were extracted from the coronal plane in $64 \times 64 \times n$ patches, where $n$ is equal to the number of input channels, and were augmented by 90 degree rotations. This resulted in between 1000–2000 training samples, depending on the original volume size, with 5% held for validation. The "Adam" optimiser was used for training, with the Dice-coefficient (Equation (6.4)) used as the loss function as described in [114], and the

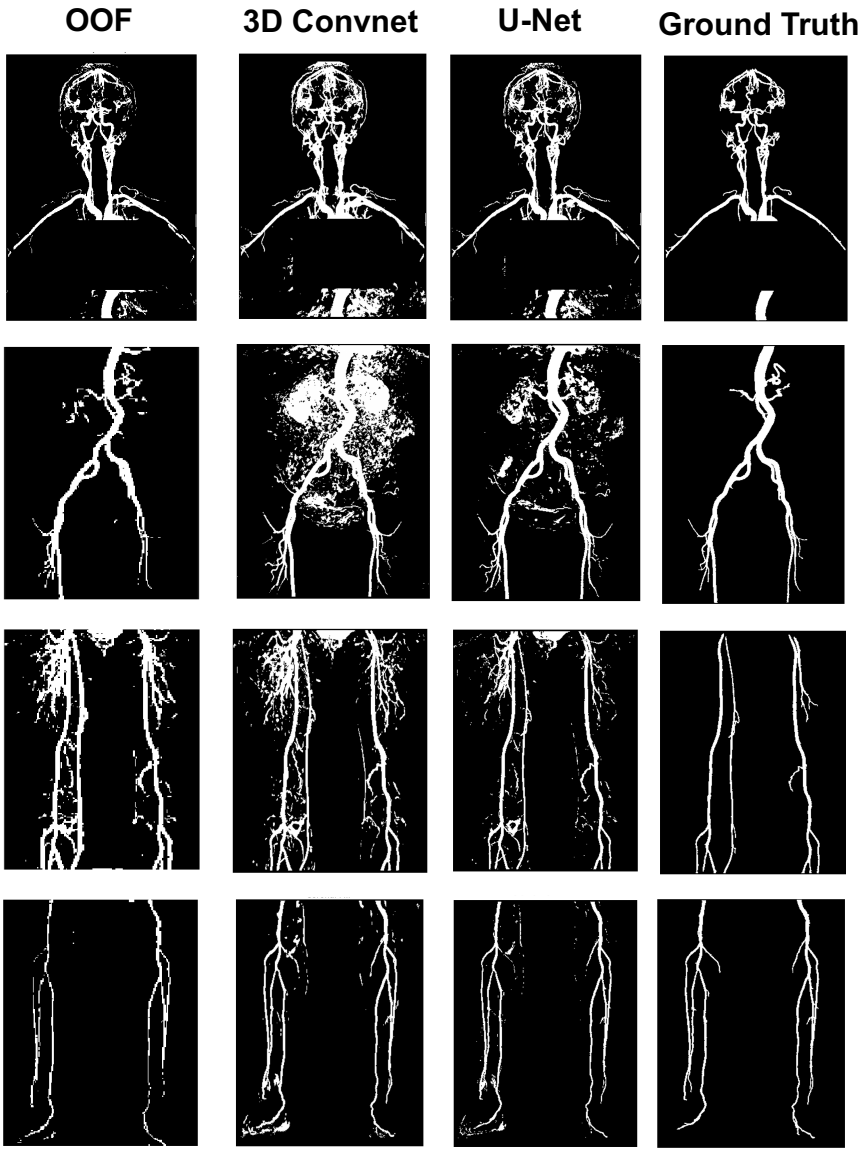



**Figure 6.10** – Coronal MIP projections of the segmentation maps for patient 1, calculated using the OOF, 3D Convnet, and U-Net methods.

best network weights determined to be those which gave the highest performance on the validation set after 100 epochs.

### 6.6.4 Results

Table 6.4 shows the results of our U-Net network with a single input channel as compared to the four previously tested techniques. We can see that the U-Net achieves the top score in 9 out of the 12 station volumes, outperforming the 3D Convnet on all stations but one despite operating on 2D slices rather than 3D subvolumes. The U-Net also achieves the best overall performance across the full dataset, with a mean Dice score of 0.756 compared to the OOF mean Dice of 0.705.

| Patient (Station) | Level Set | Frangi | OOF | 3D Convnet | U-Net |
|---|---|---|---|---|---|
| 1 (1) | 0.785 | 0.697 | 0.795 | 0.727 | **0.857** |
| 1 (2) | **0.837** | 0.817 | 0.812 | 0.618 | 0.821 |
| 1 (3) | 0.647 | 0.702 | 0.657 | 0.605 | **0.707** |
| 1 (4) | 0.672 | 0.780 | 0.843 | 0.806 | **0.870** |
| 2 (1) | 0.803 | 0.689 | 0.794 | 0.752 | **0.858** |
| 2 (2) | 0.712 | 0.722 | 0.674 | 0.552 | **0.781** |
| 2 (3) | 0.353 | 0.323 | 0.281 | 0.303 | **0.401** |
| 2 (4) | 0.662 | 0.776 | 0.844 | 0.692 | **0.883** |
| 3 (1) | 0.572 | 0.665 | 0.663 | 0.567 | **0.760** |
| 3 (2) | 0.504 | 0.488 | 0.469 | 0.422 | **0.571** |
| 3 (3) | 0.803 | 0.747 | 0.835 | **0.845** | 0.815 |
| 3 (4) | 0.618 | **0.843** | 0.791 | 0.697 | 0.742 |
| Mean DSC | 0.660 | 0.690 | 0.705 | 0.632 | **0.756** |

**Table 6.4** – Comparison of Dice coefficients for all five segmentation methods, with the highest DSC for each station highlighted in bold.

Figure 6.10 shows the coronal MIP projections of the segmentation maps for patient 1, calculated using the OOF, 3D Convnet, and U-Net methods. It can be observed that the U-Net performs much better at rejecting non-arterial artefacts than the 3D Convnet, particularly in station 2, while exhibiting better vessel continuity when compared to the OOF method.

To investigate the effect of adding 3D context to the network, we processed the data again using the same framework and training parameters, but with three input channels containing three consecutive slices.

| Input Dim. | Patient | Station 1 | Station 2 | Station 3 | Station 4 | Mean DSC | Overall Mean |
|---|---|---|---|---|---|---|---|
| (64,64,1) | 1 | 0.857 | 0.821 | 0.707 | 0.870 | 0.814 | 0.756 |
| | 2 | 0.858 | 0.781 | 0.401 | 0.883 | 0.731 | |
| | 3 | 0.760 | 0.571 | 0.815 | 0.742 | 0.722 | |
| (64,64,3) | 1 | 0.868 | 0.817 | 0.784 | 0.890 | 0.840 | 0.755 |
| | 2 | 0.853 | 0.659 | 0.409 | 0.889 | 0.703 | |
| | 3 | 0.799 | 0.527 | 0.793 | 0.774 | 0.723 | |

**Table 6.5** – Dice Coefficients for 1-channel and 3-channel inputs into our U-Net

The results are shown in Table 6.5. It can be seen that the performance of the network is almost identical despite the additional data in each patch, with a mean Dice score of 0.756 for the 1-channel and 0.755 for the 3-channel input.

### 6.6.5 Conclusions and Discussion

As an extension to the first segmentation study, we trained and tested a second CNN architecture to the same data, following the fully convolutional U-Net architecture described in [58]. Despite being trained with 2D slices, the network outperformed all previous approaches, attaining an overall mean Dice score of 0.756. Adding two additional slices to the input data by generating $64 \times 64 \times 3$ training patches did not improve the performance of the network. A recent study by Pinheiro *et al.* [115] found similar results when comparing 2D U-Net and 3D V-Net networks for brain lesion segmentation in MRI. The V-Net architecture [114] is an extension of the U-Net design for volumetric data by using 3D kernels for each convolution step. However, it was found in [115] that such 3D architectures require much more computational power than 2D and produced no significant gain in the Dice coefficient in their experiments, with the U-Net always outperforming the V-Net despite the underlying 3D nature of their data.

As a further extension to the work presented here, we would recommend a thorough comparison of U-Net and V-Net architectures, alongside more complex data augmentation techniques. In addition, supplying the output of the OOF algorithm as an additional input into a CNN may also improve performance in the presence of limited ground truth data. This could be done as a single additional channel with the maximum OOF response, or as multiple channels containing the OOF response at each radius step.

# Chapter 7

# Ground Truth Stenosis Assessments and Variability Analysis

## 7.1 Introduction

An essential step in evaluating potential algorithms for automated stenosis detection is to have ground truth data on the location and severity of stenoses in our patient scans, i.e. annotations by trained experts against which the results of the automatic algorithms will be assessed. The best performance that any system can meaningfully achieve is that its results "look like a rater's", that is, the difference between the performance (however quantified) of the system and that of any rater is comparable to the inter-rater difference. The reliability of ground truth data can therefore have a significant effect on the performance of the developed algorithms and their subsequent evaluation [116].

There currently exists no public databases of ground truth stenosis assessments for WBMRA, so this chapter covers the data collection task that we undertook in order to gather the necessary ground truth for our patient scans. There has also been no published studies of the variability and repeatability of human assessment of stenoses in MRA data. The purpose of this study was therefore twofold:

1. To gather manual assessment of stenosis locations and severities for all of our 18 patient datasets.

2. To evaluate the inter- and intra-rater variability of stenosis detection and classification from WBMRA examinations by trained observers, assessing how this may affect the development and vali-

dation of automated tools, and arriving at recommendations of how one generates reliable ground truth for this task.

The stenosis annotations analysed in this study were acquired following a clear protocol, with 18 patient datasets annotated by three clinical radiologists (one accredited radiologist, one diagnostic radiologist in training, and one interventional radiologist in training).

We will begin by examining some related studies in Section 7.2 which highlight some of the difficulties of acquiring reliable data of this type, and that there exists evidence in other studies of potentially high variability of human assessment that has not been adequately explored in the literature. Section 7.3 covers the annotation tool that was used, and the protocol that was developed for all three raters to follow. We analyse the variability of the results in Section 7.4, followed by an expert review of the high disagreement outliers in Section 7.5. Finally, we reanalyse the variability after the expert review is taken into account in Section 7.5.1 and conclude by discussing the implications of our results for the development of stenosis detection algorithms aimed at assisting or automating this diagnostic process.

## 7.2 Related Studies

The importance of validation — the process of showing that an algorithm performs correctly by comparing its output with a reference standard — in the context of medical image analysis is well established. A number of good-practice proposals have been put forward [117, 118], such as the framework for such reproducibility studies proposed by Jannin et al. [117]. In this framework a validation hypothesis is first specified with regards to expected clinical outcomes. The results of a given method are then compared against a gold-standard reference which is treated as the ground truth and the comparison results tested against the validation hypothesis, giving the validation result. In many cases, such as ours, such a gold-standard reference is either unavailable or is too difficult or costly to obtain with the next-best approach commonly adopted being to gather annotations from multiple trained annotators and then combined for increased reliability when used as ground truth for machine learning algorithms [119].

Generating sufficient ground truth for a reliable validation of machine learning algorithms for medical applications highlights a limitation of these methods in relying on the availability of ground truth. A move in

other areas of medical image analysis to leverage crowd-sourcing for annotations may be one solution but would have to achieve high reliability to meet the definition of "ground truth" [37, 38, 40, 41]. The suitability of crowd-sourced annotations remains an open question and depends on the application. We also notice recent work on the automatic generation of annotations (auto-annotations) for non-medical classifiers with large numbers of classes [42], and the growing interest of the medical image analysis community for techniques allowing to reduce drastically the number of annotations required at a parity of performance [43].

For radiological techniques, reproducibility studies are often concerned with inter-study and inter-technique variability for clinical purposes [120–122]. However recent comprehensive studies of segmentation in radiological images found that segmentation volume variability for a large group of radiologist-validated manual delineations is wide and differs significantly between structures [100], with a mean variability of up to 56%. It was concluded that "two and even three observers may therefore not be sufficient to establish the full range of inter-observer variability for this type of data, and therefore establish a reliable reference standard for the evaluation of automatic segmentation algorithms" [36].

For stenosis assessment, an early study using x-ray angiography by Fisher et al. [123] found that "When one angiographer reads a stenosis of 50% or more in the left main coronary artery, it is estimated that a second reader will report no lesion 18.6% of the time.", highlighting the variability between observers even when detecting hemodynamically significant stenoses.

A study of the reproducibility of carotid atherosclerotic plaque using MRA was carried out by Saam et al [124]. In this study the ground truth was defined using ultrasound, and assessment of the lumen, wall, and total vessel areas were used to determine modified American Heart Association lesion type (AHA-LT) on cross-sectional images. It was found that "the use of different readers substantially increased variability", with intra-reader coefficients of variation (CV) between 2.4–6.1% and inter-reader CV between 4.0–7.4%. The result was a recommendation that the same platform and the same reader be used for scans of individual subjects undergoing serial assessment, without any discussion of the implications of such variability on the reliability of the assessments themselves.

Another study by Planken et al. [125] assessed the variability of stenosis detection in forearm hemodialysis arteriovenous fistulae by multiphase

contrast-enhanced magnetic resonance angiography (CE-MRA). 15 patients were examined by 2 observers using a 5-point stenosis scale, with digital subtraction angiography (DSA) used as the gold-standard reference. Under these conditions inter-observer agreement for detection of >= 50% stenoses was just 0.81. The results also showed an instance where a vessel was assessed by one observer as having a stenosis of 75–99% (above the hemodynamically significant 50% threshold) and by the other as 0–20%. No further investigations were carried out to determine why, and under what circumstances, such disparities of opinion may occur.

A study by Naguib et al. [126] assessed a number of factors relating to the quality and reliability of a contrast-enhanced whole-body-MRA system at 1.5T, including the assessment of stenoses. Here 18 patients were assessed by 2 radiologists using a different 5-point grading scale. A total of 534 arteries (60.00%) showed no stenosis, with 78 arteries (8.76%) showing a stenosis of <= 50%, 46 (5.17%) showing a stenosis of 51–75%, 39 (4.38%) showing a stenosis of 76–99%, and 76 (8.54%) rated as occluded. A total of 117 (13.15%) arteries were deemed not assessable. In only one case of occlusion was another reference standard (DSA) used to determine the underlying ground truth. Inter-observer variability was only examined through the incorrect application of the standard Cohen's Kappa statistic to this ordinal data (which assumes the data to be categorical, i.e. unordered), split over 7 anatomical regions, where this gave a range of $\kappa = 0.75 - 1.0$. This result was deemed to show "excellent agreement between the two radiologists regarding stenosis detection and grading in all levels, except the level of the pelvic arteries and the lower leg region where the agreement was graded as good", and no further discussion of the implications of this variability was given.

It is within the above background that our study was designed, aiming to examine the whole-body MRA scans of 18 patients with 3 trained radiologists, producing a detailed statistical analysis of the intra- and inter-rater variability for stenosis severity estimation and the resulting implications for algorithm validation.

## 7.3 Annotation Tool and Protocol

A bespoke annotation plugin called the "Vessel Point Annotation" (VPA) tool for the company-confidential image visualisation software "Research Workstation" (RWS) was used for the annotation task, provided by

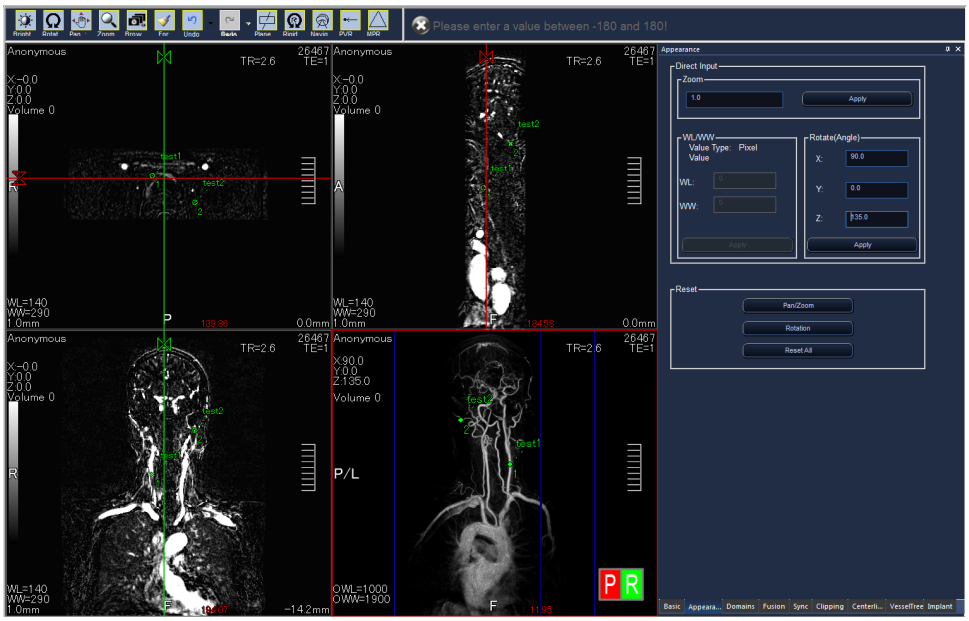



**Figure 7.1** – Main viewport of the RWS software, showing the navigation toolbar at the top left, and the 3 anatomical views and 3D rendering in the main windows.

| Stenosis Grade | Lumen Diameter Reduction |
|---|---|
| 0 | 0% (Healthy) |
| 1 | 1 – 29% |
| 2 | 30 – 49% |
| 3 | 50 – 69% |
| 4 | 70 – 99% |
| 5 | 100% (Occlusion) |

**Table 7.1** – Stenosis grade scale used during the vessel assessment.

Canon Medical Systems (Canon Medical Research Europe Ltd., Edinburgh, UK). The software displays the standard anatomical planes in three viewports (sagittal, coronal, transverse) which can be navigated independently (Figure 7.1). A crosshair in each viewport can be placed at the desired annotation location, and a marker saved with a given label selected from a predefined list. For our study this label provided the vessel name and stenosis severity, along with the exact 3-dimensional coordinates of the marker in the patient coordinate space.

For the purposes of this study we asked the raters to grade the severity of located stenosis on the scale shown in Table 7.1, based on a scale used in previous clinical studies [3, 127]. We selected 37 arteries in total (including left and right, where applicable) to be examined, based on

| Arteries Selected for Examination | | |
|---|---|---|
| 01 External Carotid (R) | 14 Abdominal Aorta | 27 Profunda Femoris (L) |
| 02 External Carotid (L) | 15 Coeliac Trunk | 28 Superficial Femoral (R) |
| 03 Internal Carotid (R) | 16 Superior Mesenteric | 29 Superficial Femoral (L) |
| 04 Internal Carotid (L) | 17 Renal (R) | 30 Popliteal (R) |
| 05 Common Carotid (R) | 18 Renal (L) | 31 Popliteal (L) |
| 06 Common Carotid (L) | 19 Inferior Mesenteric | 32 Anterior Tibial (R) |
| 07 Vertebral (R) | 20 Common Iliac (R) | 33 Anterior Tibial (L) |
| 08 Vertebral (L) | 21 Common Iliac (L) | 34 Peroneal (R) |
| 09 Brachiocephalic | 22 External Iliac (R) | 35 Peroneal (L) |
| 10 Subclavian (R) | 23 External Iliac (L) | 36 Posterior Tibial (R) |
| 11 Subclavian (L) | 24 Common Femoral (R) | 37 Posterior Tibial (L) |
| 12 Aortic Arch | 25 Common Femoral (L) | |
| 13 Thoracic Aorta | 26 Profunda Femoris (R) | |

**Table 7.2** – Arteries identified for examination based on their clinical relevance and visibility in our datasets.

their clinical relevance and their visibility in our datasets. These are summarised in Table 7.2.

The key details of our annotation protocol are summarised below. Created in close collaboration with the clinicians, it was adopted to ensure a consistent labelling procedure between raters:

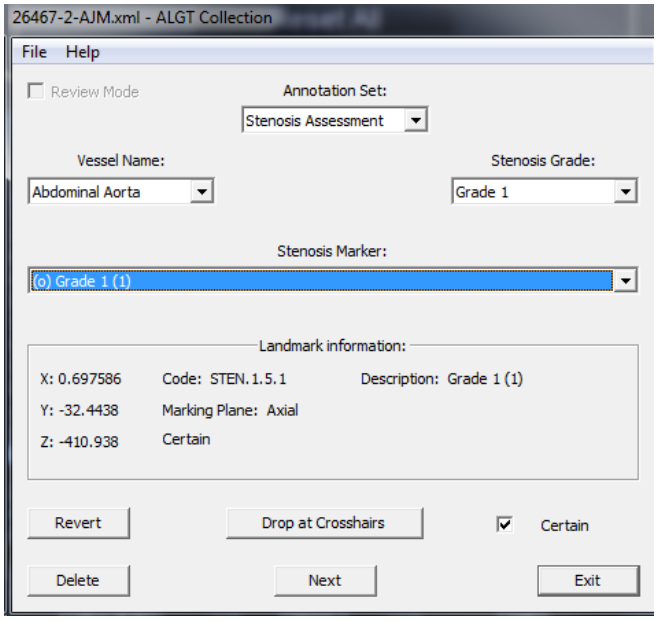



**Figure 7.2** – The VPA tool window, used to select details of the label to be placed at designated location.

**Annotation Workflow** — Each station for a single patient was to be examined sequentially, examining all arteries from Table 7.2 in each anatomical plane. For each artery the position and severity of all

found stenoses were to be marked using the annotation tool (Figure 7.2), with the stenosis severity judged by eye and assigned a number from the grading scale shown in Table 7.1.

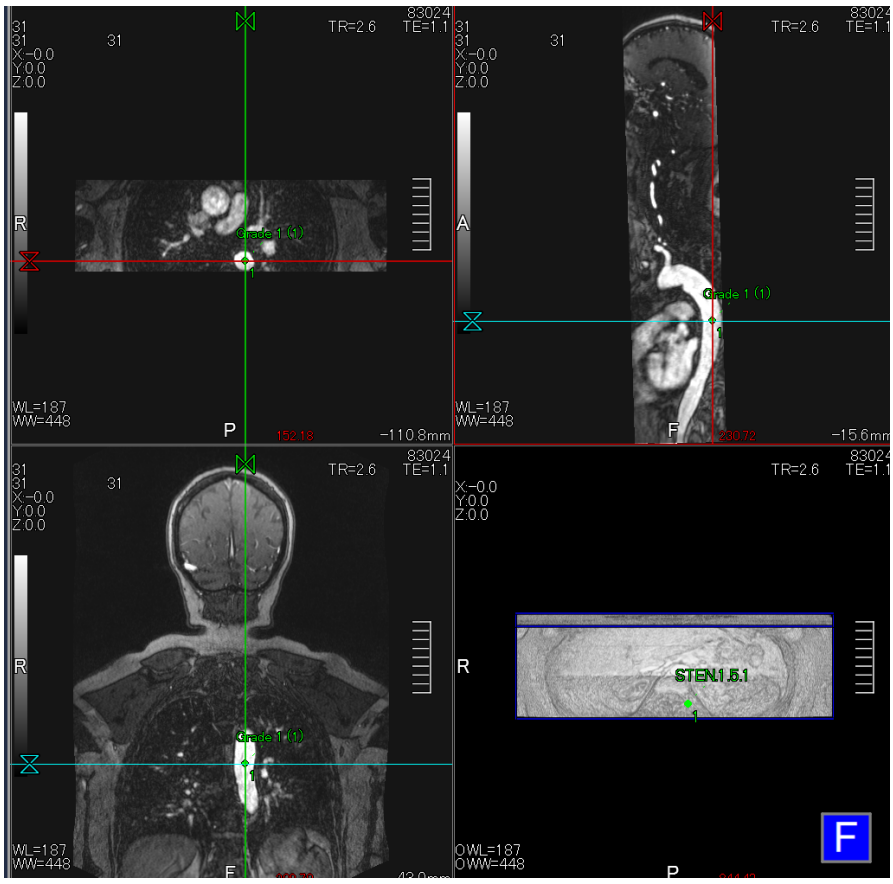



**Figure 7.3** – Localising stenosis in each view and assigning Grade marker.

**Annotating Localised Stenosis** — The crosshairs of the annotation tool were localised in the arterial lumen at the point of maximum stenosis in each of the orthogonal views, and a marker placed with the corresponding vessel label and stenosis grade (Figure 7.3).

**Stenoses in Overlapping Regions** — Since data from the individual stations was annotated separately, a stenosis may occur in the overlap regions between two stations. We asked that all stenoses are annotated in each volume independently, regardless of whether they appear in multiple stations.

**Annotating Healthy Section of Vessel** — Raters were asked to identify a single healthy region within each vessel wherever possible, in order to give us a reference for how they would expect the vessel to look in the absence of disease. This was done by placing two

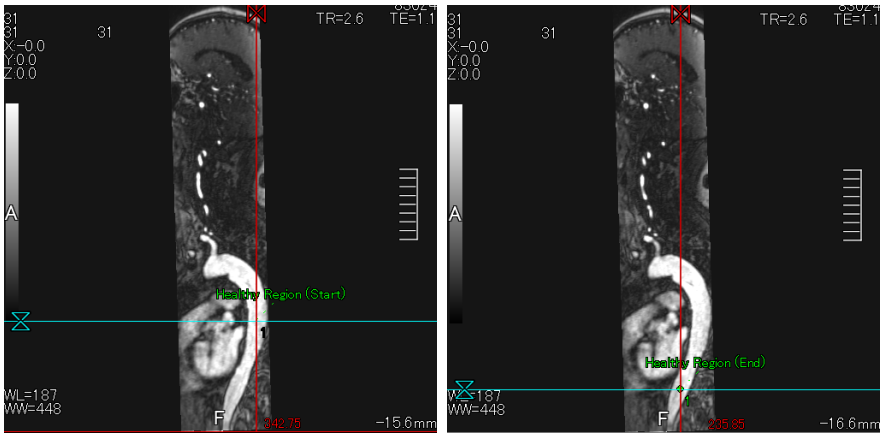



**Figure 7.4** – Annotating a healthy region.

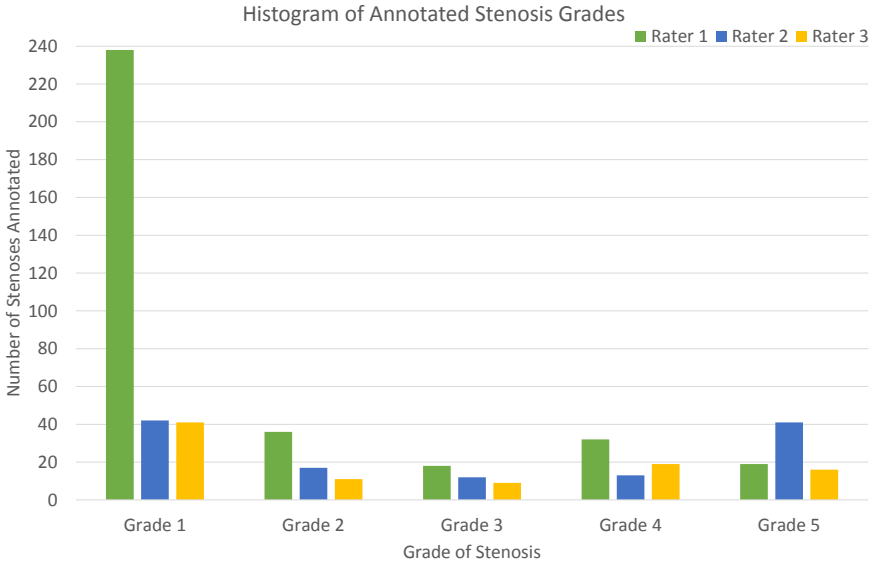



**Figure 7.5** – Total number of stenoses found per grade by each rater, for all 18 patients.

markers indicating the beginning and end of a vessel region which they considered to be normal (Figure 7.4).

**Indicating Abnormal Regions** — In cases where a vessel shows abnormalities which are not stenoses (diffuse disease where a single point stenosis cannot be determined, aneurysms, etc.), this could be indicated using two markers in the Grade menu, following the same procedure as with the healthy sections. Multiple abnormal regions could be defined in a single vessel if needed.

## 7.4 Annotation Results

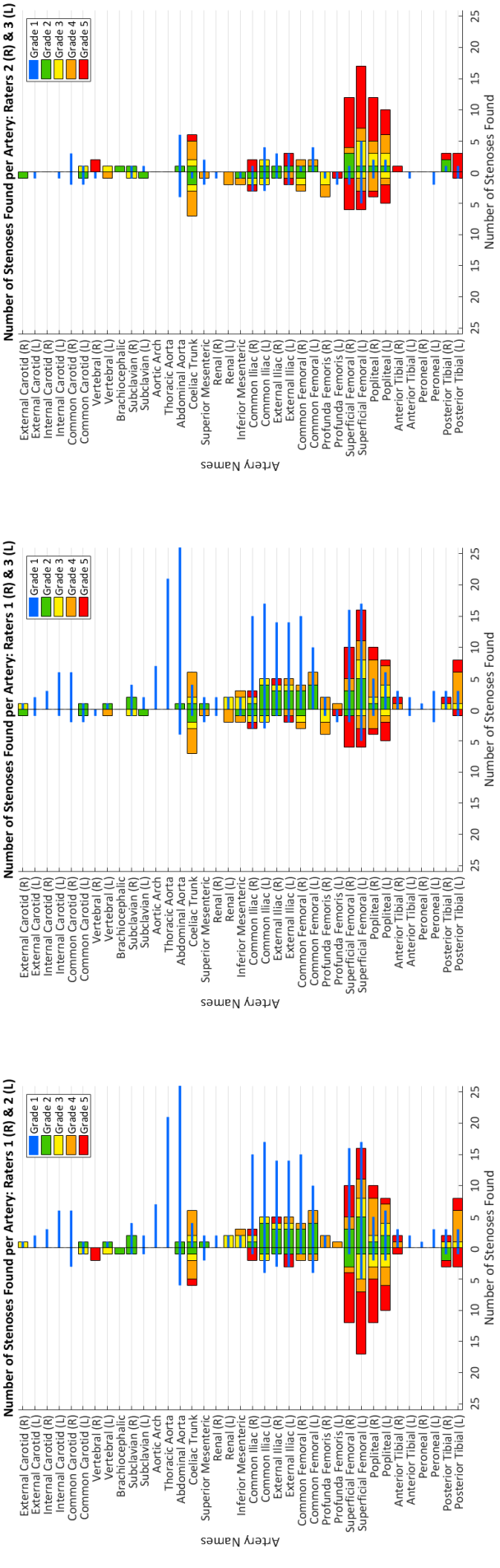
In total, 666 arteries were examined across 18 patients by all 3 clinicians. The histogram of all stenoses annotated is shown in Figure 7.5. We

| Rater Group | Percentage Agreement | Percentage Agreement (C) | Spearman Corr | Spearman Corr (C) | Krippendorff Alpha | Krippendorff Alpha (C) |
|---|---|---|---|---|---|---|
| **Six Point Scale** | | | | | | |
| 1 & 2 | 73.9 | 76.6 | 0.462 | 0.566 | 0.314 | 0.432 |
| 1 & 3 | 72.7 | 76.3 | 0.478 | 0.570 | 0.374 | 0.481 |
| 2 & 3 | 88.6 | 90.7 | 0.502 | 0.704 | 0.496 | 0.697 |
| All | 69.4 | 72.5 | - | - | 0.384 | 0.518 |
| **Three Point Scale** | | | | | | |
| 1 & 2 | 93.3 | 96.6 | 0.615 | 0.819 | 0.590 | 0.814 |
| 1 & 3 | 91.7 | 96.1 | 0.563 | 0.809 | 0.542 | 0.803 |
| 2 & 3 | 94.7 | 97.8 | 0.617 | 0.892 | 0.618 | 0.892 |
| All | 90.1 | 95.2 | - | - | 0.580 | 0.834 |
| **Binary Scale** | | | | | | |
| 1 & 2 | 95.8 | 99.4 | 0.631 | 0.946 | 0.619 | 0.946 |
| 1 & 3 | 95.5 | 99.6 | 0.609 | 0.959 | 0.601 | 0.959 |
| 2 & 3 | 96.7 | 99.3 | 0.650 | 0.933 | 0.650 | 0.933 |
| All | 94.0 | 99.1 | - | - | 0.622 | 0.946 |

**Table 7.3** – Agreement, correlation, and reliability measures for different rater groups. Columns marked (C) represent the numbers after the outlier disagreements were corrected by the 4th rater. The upper set uses the original grading scales, with the other two calculated using a redefined number of grades; a 3 grade scale (combining 0+1, 2+3, and 4+5 grades), and a binary scale with the threshold set at 50% stenosis. In all cases, $p < 0.005$.

can see significant differences between raters across all grades. In grades 2–5 there is a 2–3 times difference between the maximum and minimum number of stenoses found, with the largest differences found in the Grade 1 category (< 30% diameter reduction); while raters 2 and 3 found similar numbers of 41 and 42 respectively rater 1 found 238, almost 6 times as many. These low grade stenoses can be very subtle and are typically not recorded in a diagnostic setting. The large disparity in numbers can therefore be attributed to the different backgrounds of each rater, where rater 1 had more experience with identifying low grade stenoses as part of previous research studies.

Figure 7.6 shows the distribution of stenosis grades throughout the body, from which a number of observations can be made (note that the numbers of grade 1 stenoses varies substantially between raters, hence they were plotted as a separate bar). In general it was found that higher grade stenoses occur in patients with a larger number of overall stenoses (due to more advanced disease), and the majority of high grade stenoses occur below the aortic bifurcation. We can also see that there is a consis-




(a) Pair 1 (b) Pair 2 (c) Pair 3

**Figure 7.6** – Comparison of annotated stenosis distributions, showing the number of stenoses per grade per vessel, arranged vertically, superior to inferior.

tent trend between raters of the highest concentration of $> 30\%$ stenoses being located towards the extremities in the femoral, popliteal, and posterior tibial arteries, with an additional spike in the coeliac trunk, which is consistent with the expected distribution in peripheral arterial disease patients [128].

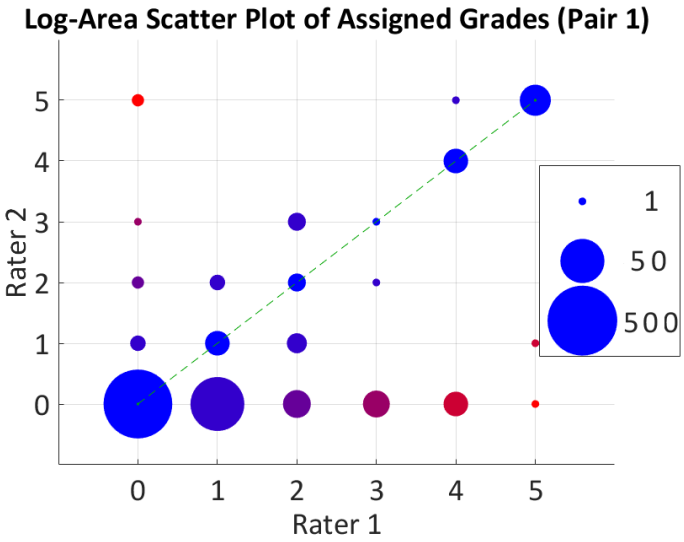
The comparative statistics for each rater pair is given in the first section of Table 7.3. These were calculated by assigning each vessel a single grade based on the maximum stenosis present for each rater, and calculating the statistics on those 666 artery grades. We can see that we achieved an overall agreement of 69.4%, with 15% higher agreement between raters 2 and 3 as compared to the pairs with rater 1. The average Spearman correlation coefficient is 0.481, and the Krippendorff's alpha coefficient (valid for multi-class, ordinal, multi-rater data) is 0.384. A general interpretation of this statistic is $\alpha >= 0.8$ gives high reliability, $0.8 > \alpha >= 0.667$ gives low reliability, and data for which $\alpha < 0.667$ is unreliable [129]. We can therefore see that, taken as a whole, our combined annotation pool using the 6 point grading scale in Table 7.1 is not reliable.

Table 7.4 shows the intra-rater variability, calculated from 3 patient datasets which each rater examined a second time at the end of the initial study. To avoid memory effects, the patients were chosen from early in the study and examined around 3 – 4 months after their initial examination, with the raters blinded to their first set of annotations. We can see that rater 1 is the most consistent with a Krippendorf's Alpha of 0.882, followed by rater 2 at 0.796 and rater 3 at 0.701. The percentage agreement is between 80 – 82% on the six point scale, rising to 89 – 92% for the three point scale. These lower than expected repeatability results demonstrate the complexity of the annotation task being undertaken, and need for consensus between multiple raters when formulating ground truth assessments.

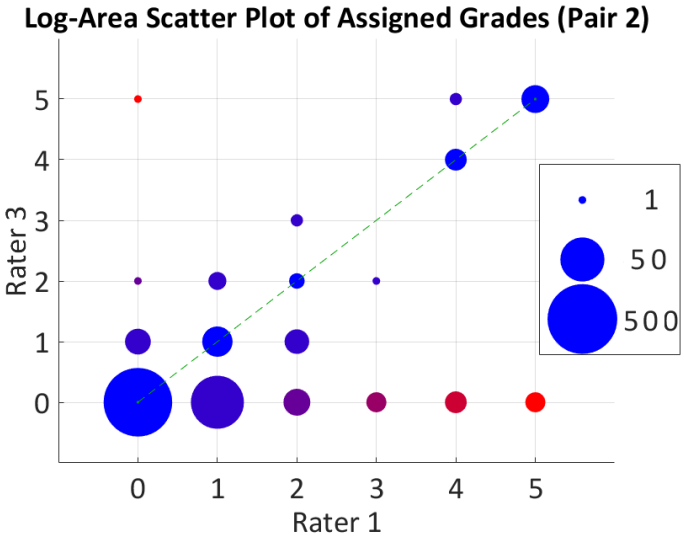
Finally, the disagreement between rater pairs is shown graphically in the scatter plots of Figure 7.7, with the exact numbers given in the inter-rater agreement matrices in Figure 7.8. This was again constructed using the maximum stenosis grade per vessel, with the area of each circle representing the number of cases on a logarithmic scale. Given the natural difference of opinion expected when distinguishing between adjacent grades by eye, we would expect the data to be clustered around the line of agreement by $\pm 1$ grade. However we can see that there are a number of very high disagreement cases; for example for pair 2 there are 6 cases

| Rater | Percentage Agreement | Spearman Correlation | Krippendorff's Alpha |
|---|---|---|---|
| **Six Point Scale** | | | |
| 1 | 80.2 | 0.884 | 0.882 |
| 2 | 82.0 | 0.797 | 0.796 |
| 3 | 82.0 | 0.700 | 0.701 |
| **Three Point Scale** | | | |
| 1 | 91.9 | 0.861 | 0.862 |
| 2 | 89.2 | 0.766 | 0.764 |
| 3 | 91.0 | 0.721 | 0.722 |
| **Binary Scale** | | | |
| 1 | 96.4 | 0.888 | 0.887 |
| 2 | 95.5 | 0.818 | 0.813 |
| 3 | 94.6 | 0.794 | 0.793 |

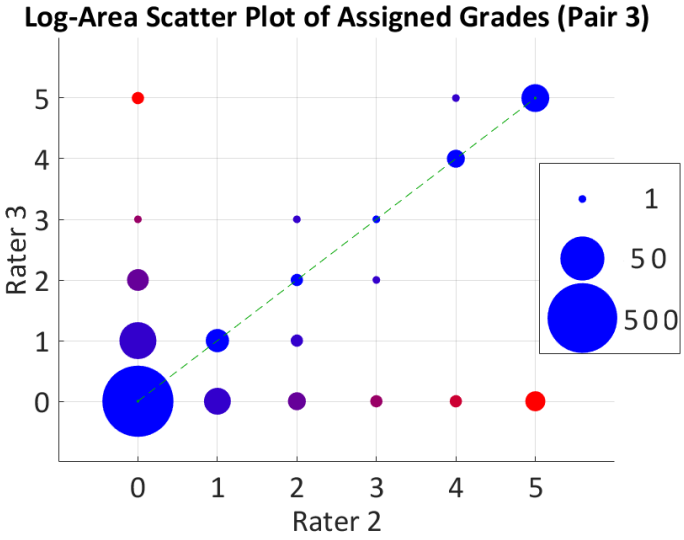
**Table 7.4** – Intra-rater agreement, correlation, and reliability measures. The upper set uses the original grading scales, with the other two calculated using a redefined number of grades; a 3 grade scale (combining 0+1, 2+3, and 4+5 grades), and a binary scale with the threshold set at 50% stenosis. In all cases, $p < 0.005$.




(a) Pair 1

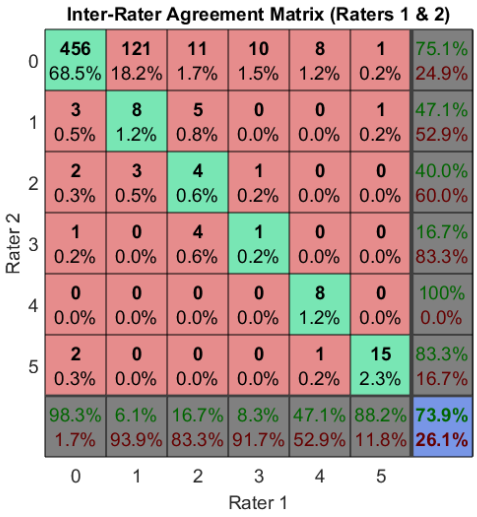



(b) Pair 2

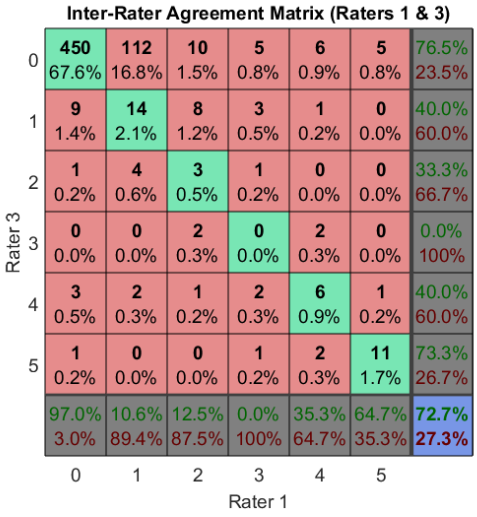



(c) Pair 3

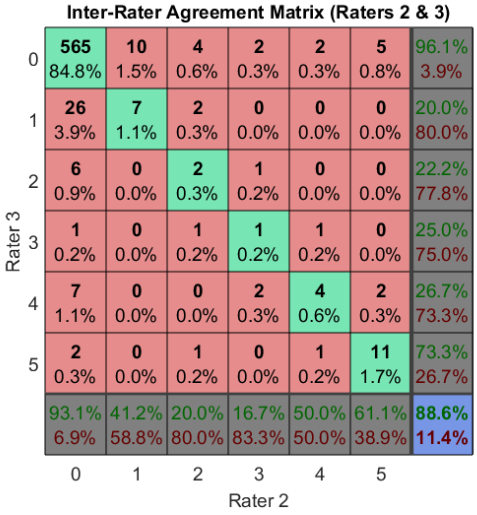
**Figure 7.7** – Scatter plots of the maximum assigned grade per artery for all vessels across all patients, where the area of each dot represents the number of cases on a logarithmic scale.




(a) Pair 1




(b) Pair 2




(c) Pair 3

**Figure 7.8** – Inter-rater agreement matrices for each annotator pair, showing the exact numbers of each possible pairing which is shown graphically in Figure 7.7.

| Min. Difference \ Min. Stenosis | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 204 | 72 | 45 | 31 | 13 |
| 2 | 58 | 58 | 37 | 26 | 10 |
| 3 | 36 | 36 | 36 | 25 | 10 |

**Table 7.5** – Number of disagreement cases, given the minimum severity of stenosis involved, and the minimum number of grades between the highest and lowest assessment.

where one rater has assigned a grade 5 stenosis, and the other has said the vessel is healthy (grade 0). Given the unexpected spread of the data, an additional review task was undertaken to provide a final vessel score and explore the possible reasons for these high disagreement cases.

## 7.5 Expert Review of High Disagreement Cases

A review procedure was undertaken with a consultant radiologist with more than 20 years experience in MRA analysis. The task was to review arteries which contained high levels of disagreement between at least two raters, giving a final authoritative assessment of these cases. These arteries were determined by first assigning each artery a single grade based on the maximum stenosis present and then calculating the differences in grades between raters. Table 7.5 shows the number of disagreement cases tabulated by the minimum severity of stenosis involved, and the minimum number of grades between the highest and lowest assessment. We selected all examples containing a stenosis assessment of at least Grade 3, on which at least one rater disagreed by at least 2 grades, as being the most severe and clinically relevant. There were 37 arteries in which such disagreements occurred.

The review procedure was carried out using the same viewing software employed during the initial assessment. For each example the original datasets for that patient were presented to the reviewer, along with the three previously assigned grades and corresponding maximum stenosis locations. Using this information the reviewer gave his own assessment of the maximum stenosis present in that artery, as well as his estimate

of why the disagreements had occurred.

It was found that the likely causes of these high disagreement cases could be grouped into eight categories given below (with the number of occurrences given in parentheses). Note that not all disagreements could be assigned a clear cause.

**User Error with the Annotation Tool (6)** These occurred where the user had obviously selected the wrong vessel label for the stenosis marker (e.g. selecting left common femoral instead or right common femoral).

**Missed Stenoses (5)** In a number of cases there were no artefacts or other obvious confounding factors which could have led to a particular stenosis being missed. These missed stenoses were therefore assumed to be observer errors.

**Differences of Opinion (4)** Some cases (particularly those in the grade 2 – 4 range) were found to be simply due to differences of opinion as to how severe the stenosis was. This mainly affected thin vessels where the resolution of the scan gave a restrictive number of samples across the artery, making accurate grading difficult.

**Image Artefacts (4)** Movement or noise was found to impact the ability to assess certain vessels, particularly near the edge of the image volumes. In some cases these could be worked around by referencing the overlap region of the next station in order to get a clearer view, but not all raters may have done this.

**Vessel Name Uncertainty (3)** It has been shown in other studies that the location at which a vessel name changes can have a high uncertainty ($> 20mm$ in many cases) [130], e.g. where the external iliac becomes the common femoral. If a stenosis occurred around these regions the same stenosis could have been assigned to different vessels by different raters.

**Poor Image Quality in Overlap Region (3)** A few stenoses were found in the overlap region between two stations. The image quality differed significantly between the two stations in this region for the cases in which high disagreements occurred. It was therefore suspected that a stenosis may have been missed if the user did not check the area in both stations.

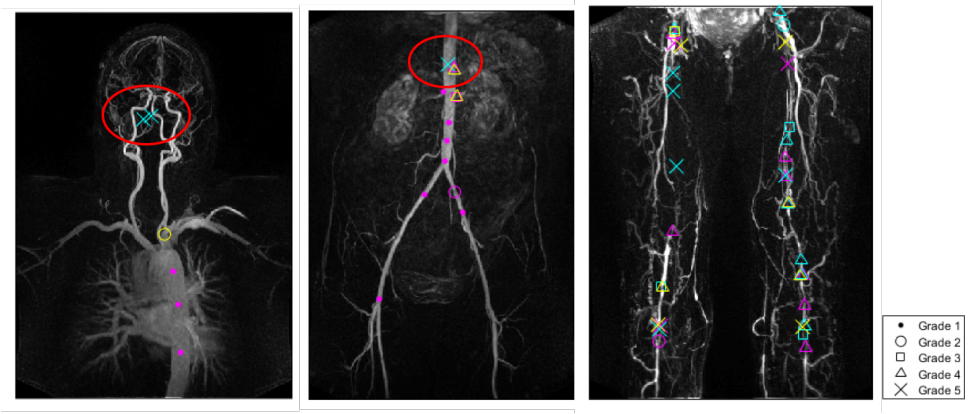



**Figure 7.9** – Maximum intensity projection with annotated stenosis marker overlays, where the severity is given by the symbol and the raters are differentiated by colour. Particular examples of disagreement cases are circled in red.

**Distal Vessels (2)** Some stenoses occurred in the most distal regions of the vascular tree, where both reduced vessel contrast and increased noise made it difficult to assess the true extent of the stenosis. The decision of where to stop assessing the vessel based on the image quality may also have therefore led to certain raters missing very distal stenoses.

**Natural Narrowing and Slender Vessels (2)** Finally, in some very slender arteries the natural narrowing which can occur may appear to create a stenosis but was assessed as healthy in the consensus.

Some examples of the above disagreement cases are shown in Figure 7.9, showing cases of natural narrowing being misclassified as pathology, differences of opinion, and ambiguity of vessel naming, particularly in areas of high disease. The complexity of the distribution of labels seen in Figure 7.9, coupled with the uncertainty in annotation position found in a similar study ($> 20mm$ in many cases [130]), meant that a positional analysis could not be easily undertaken without significant time spent with all three annotators, which was not possible for this study.

### 7.5.1 Results of Review Procedure

The histogram of the reassessed artery grades is shown in Figure 7.10. We can see that the majority of these high disagreement cases occurred at the extremes of the scale ($< 30\%$ and $\geq 70\%$), with only 4 of the 37 cases falling in the grades 2 – 3 range. The disagreement at the lowest end of the severity range may have been due to a lack of familiarisation of the raters with assessing stenoses of less than 30% which are not

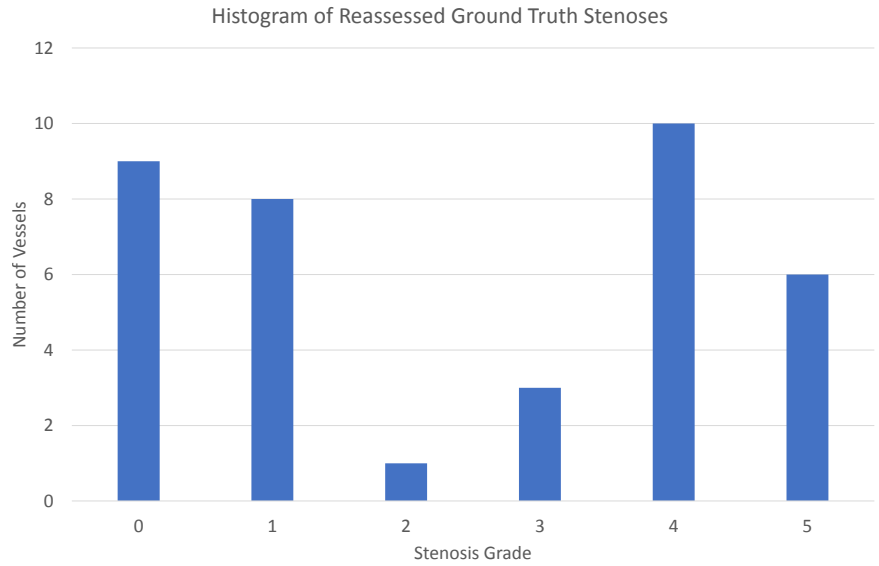



**Figure 7.10** – Histogram of grades assigned to all reassessed vessels by 4th expert rater.

| Rater | Percentage Agreement | Percentage Overesti-mated | Percentage Underestimated |
|---|---|---|---|
| 1 | 62.2 | 27.0 | 10.8 |
| 2 | 40.5 | 8.1 | 51.4 |
| 3 | 29.7 | 24.3 | 46.0 |

**Table 7.6** – Statistics for percentage agreement, overestimation, and underestimation of artery grades compared with the expert reassessment.

significant in clinical practice, while the higher number of high grade disagreements likely stemmed from the patients with advanced PAD where the assessment of both vessel name and stenosis severity becomes extremely difficult (as in the example shown in Figure 7.9).

Figure 7.11 shows the histogram of disagreements between each original rater and our 4th expert, where positive numbers indicate that rater underestimated the grade compared to the reassessment, and negative numbers show they overestimated. Firstly we can see that rater 1 agreed the most with the reassessment giving the same score 62.2% of the time, followed by rater 2 then 3 with 40.5% and 29.7% respectively (Table 7.6). We can also see that when they disagreed, rater 1 tended to overestimate the true severity, while raters 2 and 3 tended to underestimate.

Using these reassessed grades we "corrected" the data by replacing the rater's grade with the reassessment if they disagreed by 2 grades or more. The statistics for our new data pool is shown in Table 7.3 in

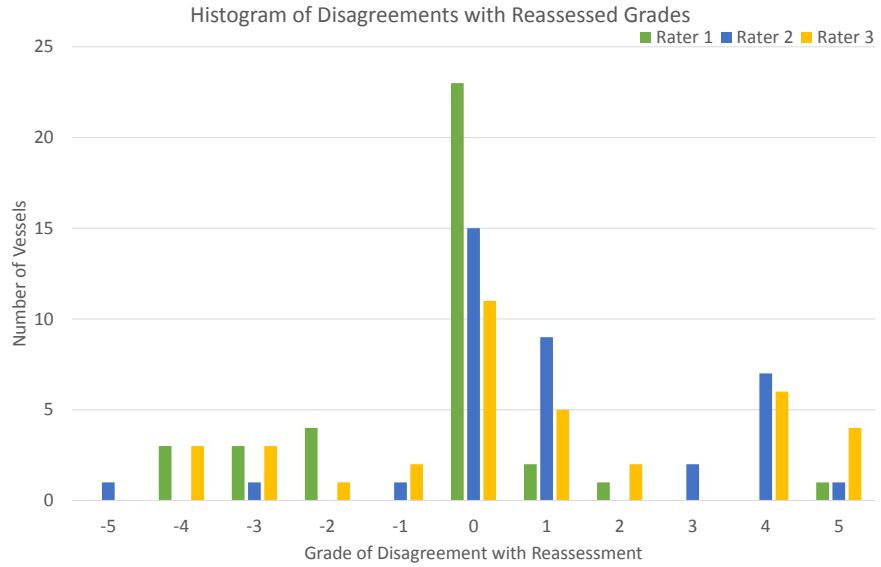



**Figure 7.11** – Histogram of disagreement with the 4th expert rater. Histogram shows reassessed grade minus the original grade, so negative numbers correspond to the original rater giving a higher severity than the expert reassessment.

the columns labelled "Corrected". We can see that this has a nominal effect on the groupings involving rater 1 where the disagreements are dominated by the grade 1 vs grade 0 disagreements (Figure 7.7). It has a much greater effect on raters 2 and 3, where a small increase in agreement (88.6% to 90.7%) leads to a large increase in the Spearman correlation from 0.502 to 0.704. However, we can see that the corrected data from all raters still has much lower reliability than desired, with just 72.5% agreement and $\alpha = 0.518$.

A further step was therefore taken, where the effect of re-binning the data into a smaller grade scale was examined. Following advice from our clinical partners we created a three point scale — combining grades $0 + 1$, $2 + 3$, and $4 + 5$ — and a binary scale, setting a threshold at grade 3 (50% stenosis). The results for these new scales are shown in the middle and bottom sections of Table 7.3. We can see that moving to a three point scale greatly increases the agreement between raters from 69.4% to 90.1% in the original dataset, and from 72.5% to 95.2% after corrections. Similarly the reliability score increases from 0.384 to 0.58, with the corrected data pushing above the desired 0.8 value with $\alpha = 0.834$. The binary scale shows an additional smaller improvement beyond this, however note that the reliability of the uncorrected data is still below 0.8 even on the binary scale.

## 7.6 Conclusions

The purpose of this study was to assess the inter- and intra-rater variability of stenosis classification from WBMRA examinations, using data acquired for the specific task of generating ground truth for software development, and evaluate how this variability may affect the development and validation of such tools.

Three clinicians of similar training experience examined 18 patients, generating data with higher variability than anticipated with just a 69.4% agreement when assessing the maximum stenosis grade per vessel, and with intra-rater agreement between 80 – 82%. We found there was a higher degree of disagreement in the low severity stenoses, particularly those below 30%, with higher concordance among the higher grades.

Undertaking a review procedure of the highest disagreement cases with a 4th expert rater significantly improved the overall agreement and reliability of our data and revealed many of the reasons why such disagreements can occur, however the reliability was still below the level at which robust conclusions could be drawn. We therefore conclude that for this task, using the experimental setup described here, a six point grading system is not suitable due to being too fine-grained for the stenosis assessment method used (assessment by eye).

It was found that a three point grading scale led to a much better consensus with our data. The agreement, correlation and reliability statistics on this scale were greatly improved, with the reliability in particular rising above 0.8 when the additional review corrections were taken into account. Further justification for such a scale can be seen by observing the arteries which contain the highest concentrations of severe stenoses. From the stenosis-per-vessel figures in Figure 7.6, the highest concentrations of high grade stenoses are in the femoral, popliteal, and posterior tibial arteries. These were found to cover a range of 2–10mm in diameter, in agreement with previous studies of similar populations; 6–10mm for the femoral [131], 5–8mm for the popliteal [132], and 2–4mm for the posterior tibial arteries [133]. Our resolution range is 0.98–1.3mm per pixel, giving these arteries of high interest an average diameter of 4–5 pixels in our data. When assessing such arteries by eye a three point grading system — “healthy”, “low severity”, “high severity” — is the best which could be realistically expected to give consistent results.

Our results demonstrate that assessing the exact location and severity of stenoses from WBMRA examinations is a complex clinical decision presenting even substantial variations among multiple trained experts.

The collection of ground truth to validate automatic stenosis detection systems therefore requires great care, particularly with regards to the development of a clear protocol and careful selection of the grading scale to be used, and should be undertaken with multiple raters as has been done in other large scale studies in other modalities [134].

# Chapter 8

# Automated Stenosis Grading

## 8.1 Introduction

In this chapter we will discuss the three algorithms which were developed for automatically analysing vessel segments and assigning a stenosis severity score. The input to each algorithm is a pre-processed volume, which has gone through the first four processing stages of the pipeline shown in Figure 1.3, and the output is the stenosis grade based on the maximum stenosis present. This approach, and the scale used, follows the form of the ground truth we collected and we will analyse the results on the same three scales which were derived in Chapter 7.

We will first introduce the three algorithms in Section 8.2, examining the technical details of each. In Section 8.3 will then apply them to synthetic data, constructed to mimic the expected form of the straightened vessel subvolumes produced by our pipeline. We then apply these algorithms to our 18 patient datasets in Section 8.3, using the ground truth data as described in Chapter 7. Finally we will conclude with a discussion of the results, analysing how they compare between the real and idealised case, and what the reasons for the differences in performance might be.

## 8.2 Stenosis Grading Algorithms

The input volume for each algorithm described here is in the form of a straightened vessel subvolume which has been tracked in 3D space using the centreline. An example of a typical tracked vessel produced by our automated processing pipeline is illustrated in Figure 8.1.

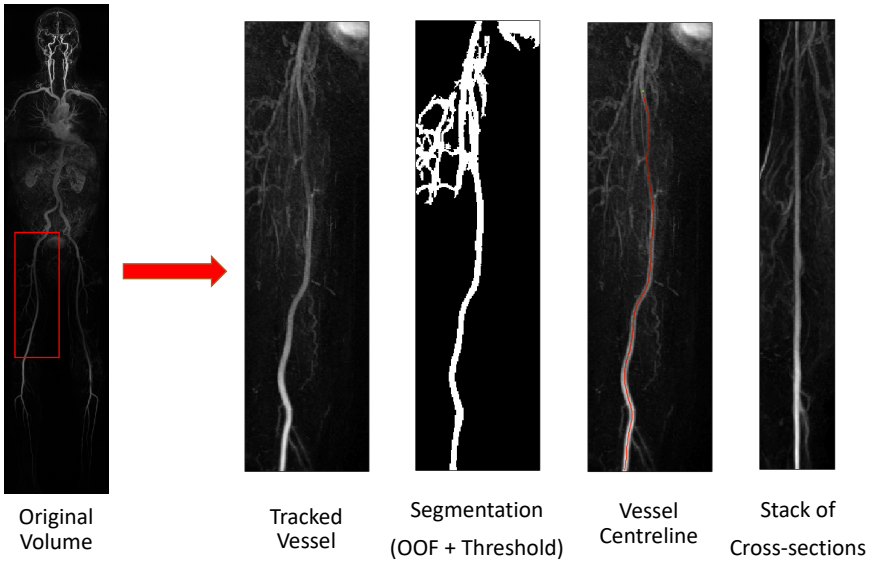



**Figure 8.1** – An illustration of the pre-processing stages of our pipeline before the stenosis grading algorithms are applied.

### 8.2.1 Cross-Sectional Area

The first algorithm is based on the cross-sectional area of the vessel, from which the equivalent diameter is calculated at regular one-voxel steps along the centreline.

The following procedure is followed for calculating the diameters for each input volume:

1. The straightened vessel subvolume and its corresponding segmentation mask are loaded.

2. At each step along the centreline — i.e. each axial slice of the straightened volume — the number of non-zero voxels in that slice of the segmentation mask is tallied.

3. The cross-sectional area for that slice is recorded as the total number of non-zero voxels at that position, and the equivalent diameter calculated by assuming a circular vessel cross-section.

4. The slice position is incremented by one along the vessel axis and the above procedure repeated.

This produces a vector of estimated diameters along the entire length of the vessel volume. Once this has been calculated, the following algorithm is used to derive the corresponding stenosis severity (as a percentage, with 0% being healthy):

1. First, an estimate of the "healthy" diameter of that vessel at each point along the vessel is estimated. This was done by first assuming that only stenoses are present (no aneurysms), and therefore the healthy parts of the vessel are its widest points. We also assume that the stenoses are small in length compared to total length of the vessel and the vessel may naturally taper towards one end.

2. A healthy estimate of the vessel diameter is derived from the linear least squares fit of the largest 50% of the diameter measures, creating an expected "healthy" value at each point along the vessel.

3. For each slice along the vessel, the difference between the calculated diameter and the estimated healthy area is calculated.

4. An estimation of the background noise for that vessel (set empirically to be two standard deviations for our data) sets the lowest threshold for which a stenosis can be detected.

5. If the largest difference between the calculated diameter and expected healthy diameter is higher than the estimated noise threshold, then a stenosis was detected, i.e. the reduction in diameter from the expected healthy value was larger than could be attributed to image noise.

6. In order to prevent the detection of narrow outliers due to noise or inaccuracies in the segmentation mask, the neighbouring points on either side of the maximum stenosis are also considered to determine if they too are above the threshold, using the assumption that a true stenosis must be elongated by at least three voxels. If all three points meet the above criteria then a stenosis is recorded.

7. If a stenosis is detected, the vessel is assigned a severity score by the formula $S_v = 1 - (A_s/A_h)$ where $S_v$ is the stenosis score for vessel $v$, $A_s$ is the diameter at the point of maximum stenosis, and $A_h$ is the estimated healthy diameter at the same location. This gives a number between $0 - 1$, with 0 being completely healthy, and 1 being a complete occlusion of the vessel.

Figure 8.2 shows an example of the measurements obtained by this algorithm, as applied to a synthetic vessel volume with an idealised stenosis of 80%.

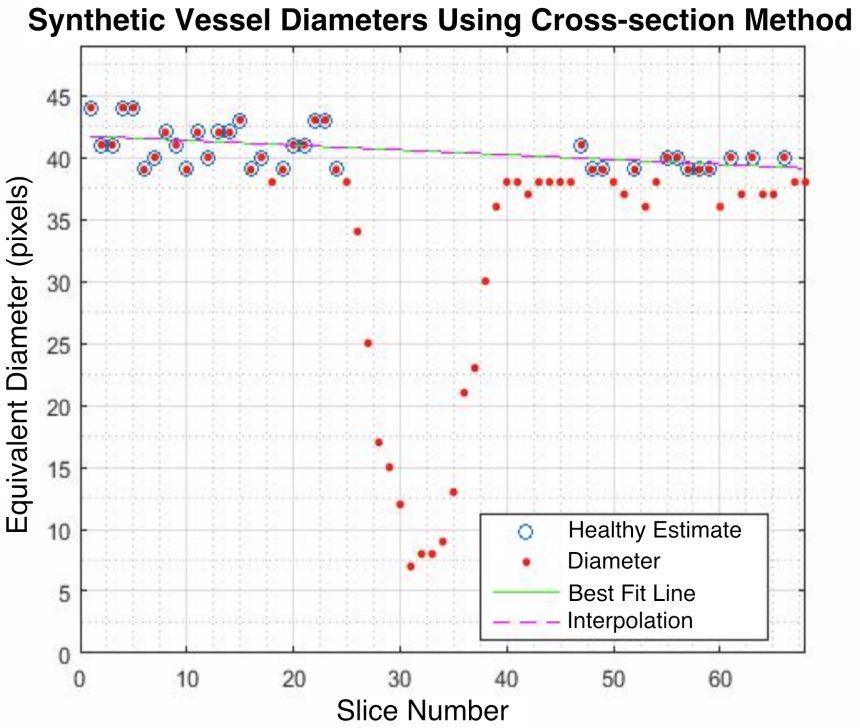



**Figure 8.2** – A graph of the calculated diameters of a vessel segment with an idealised stenosis. The points used for the healthy estimate are circled in blue, the line of best fit shown in green, and the sampled healthy estimate points at each location shown in purple.

### 8.2.2 GroBa Calibre

The second algorithm uses the GroBa method to estimate the vessel calibre, as explored in Chapter 5.3.

The algorithm follows the exact same steps as described for the cross-section method, differing only in the measurements derived at each point along the vessel. Instead of tallying the number of non-zero voxels in the segmentation mask, a "balloon" is inflated from the seed point on the centreline (the central voxel on the given slice). This is expanded into binary volume until the standard stopping criteria is reached (the balloon length is twice its width, or no more voxels can be added), and the equivalent diameter derived from the equation of a cylinder, as described in Chapter 5.3.

### 8.2.3 Convolutional Neural Network Model

The final approach was based on a convolutional neural network, motivated by the recent successes of deep learning networks for many medical imaging tasks [135].

There are no analytical methods by which the optimal structure of a network can be determined a priori for a given problem, so a number of

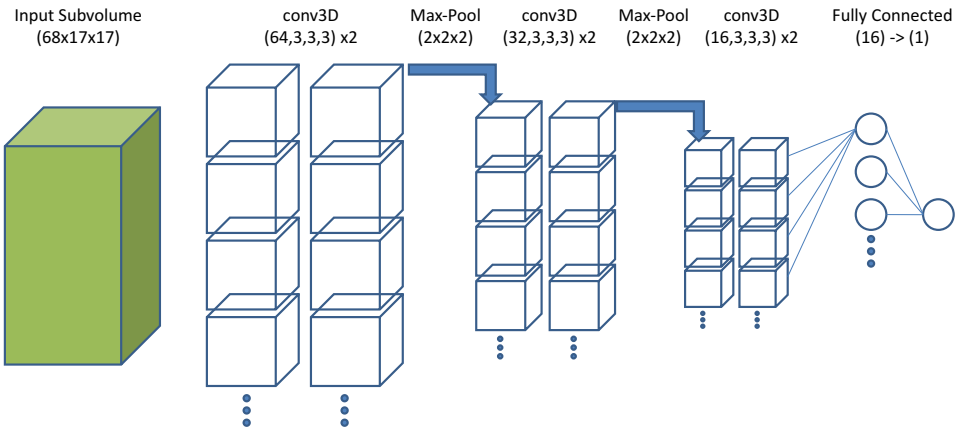



**Figure 8.3** – Schematic of the CNN used

network structures were tested in order to determine the best match for the complexity of our problem. It was found that if the network was too shallow it would give unstable performance during training and often fail to converge. Whereas deep networks were mismatched to the amount of data we have available for training, causing it to rapidly overfit the data and give very poor results in testing.

The final structure presented here was based on the best performing network when applied to the real data as described in Section 8.4. A schematic of the final network structure is shown in Fig. 8.3, which was implemented using Keras and Tensorflow (tensorflow-gpu 1.1.0). It consists of three sets of two convolutional layers followed by a max-pooling layer, which connects two a fully connected layer of 16 nodes, and finally a single fully connected node for the output severity rating.

All layer activation functions were "ReLU" except the final output node, which was a customised ReLU function which gave zero response below zero, a linear response between 0 and 1, and a saturated response of 1 for all values above 1. This ensured an output value of between 0 and 1 corresponding to the maximum percentage stenosis present. The stochastic gradient decent (SGD) optimiser was used during training, with mean squared error selected as the loss function. While various loss functions have been used with ordinal data in the literature, mean squared error is commonly used as it has been previously shown to be suitable for such data [136].

## 8.3 Stenosis Grading with Synthetic Data

We first tested our stenosis grading algorithms using synthetic data. This was done in order to investigate how well they performed in the ideal case,

where the difficulty of the problem could be controlled and the ground truth is completely accurate.

We will first discuss how the synthetic data was generated, before examining the performance of each algorithm, evaluated in the same manner as the real ground truth data in Chapter 7.

### 8.3.1 Synthetic Data Generation

There are currently no synthetic vascular datasets available for WBMRA, so we created our own from scratch. It is important to note that our target is not to create a realistic vascular tree, but rather the straightened vessel segments produced by our processing pipeline, similar to the example shown in Figure 8.1.

As such, the following assumptions were made when creating our synthetic vessel volumes:

1. The vessel has been rectified, i.e. the curved path of the vessel in 3D space has been tracked using a centreline, and a "straightened" volume produced by taking axial cross-sections at each step along the centreline.

2. The vessel intensities have been normalised, and any inhomogeneities due to MR artefacts removed.

3. The target vessel has been tracked between junction points, and therefore no branches are present in the generated vessel.

4. There are no other abnormalities present apart from localised stenoses (e.g. aneurysms).

The vessels were generated by initialising an elongated empty vessel volume, then stepping through it slice by slice and at each cross-section creating a circle of a given diameter and filling all enclosed voxels with the desired intensity. This was then smoothed with a gaussian filter and random noise added. In order to give a smooth appearance, the vessels were created at a high resolution, and then downsampled to the desired size of $68\times17\times17$. This was found to give a visually similar structure to the vessels seen in our real data.

Stenoses were added using the tapered cosine (or "Tukey") window function, adding a localised reduction in diameter along the vessel. This was used to give a visually reasonable form to the stenosis, while having easily tunable length, smoothness, and maximum severity.

The following parameters were varied randomly between each generated vessel, with the parameter ranges given in brackets which were determined empirically to reflect the true values seen in the real data:

1. Healthy diameter (4–9 voxels).
2. Natural taper along the length of the vessel (0–10%).
3. Centreline noise, i.e. displacement of the vessel centre at each slice (order of 0.5–1.5 voxels, smoothly varying along the length).
4. Sigma of gaussian smoothing applied to the vessel (0.7–0.9).
5. Noise mean and variance, applied from a uniform distribution (0.003–0.007 and 0.0001–0.0002 respectively).
6. Mean intensity of vessel (0.8–1.2).
7. Number of stenoses (0–5).
8. Severity of each stenosis (0–100%).
9. Stenosis length (8–24 voxels).
10. Stenosis Tukey alpha (0.8–0.99).
11. Stenosis location within the vessel.

In all cases, the ground truth grade was recorded as the severity of the maximum stenosis present, given as the percentage diameter reduction.

Figure 8.4 shows example MIPs of the generated vessels. The first row shows single stenoses of the listed severity, with the bottom row showing a second example with multiple stenoses, the most severe of which corresponds to the quoted severity.

### 8.3.2 Results

For our experiments we generated a set of ten thousand synthetic vessels, which were analysed by all three algorithms. The dataset was balanced to give equal numbers of vessels within the grade bins used for our real ground truth annotations. For the CNN method, the network was trained by splitting the data into 5 equal folds, with training set in each fold consisting of 8000 vessels with 5% used for validation, and then after 50 epochs the network was tested on the held-out 2000 vessels.

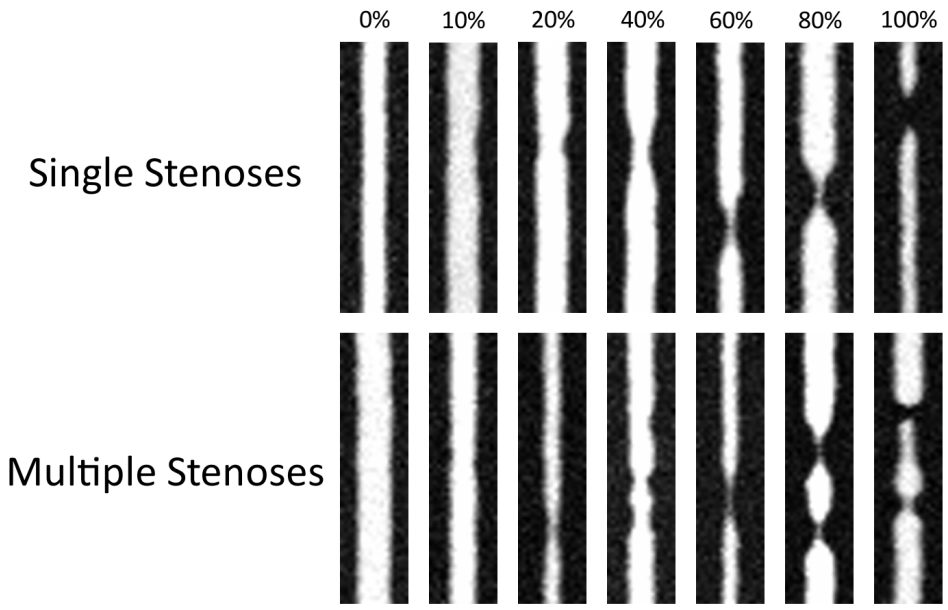



**Figure 8.4** – Example MIPs of synthetic vessel volumes, containing single (top) and up to 5 (bottom) stenoses.

Figures 8.5, 8.6 and 8.7 show the results as confusion matrices analysed at the three different scales used in Chapter 7 (6 point scale, 3 point scale, and binary).

Firstly, it can be observed that both the cross-section and GroBa methods exhibit a saturation effect whereby high-grade stenoses are often classified as occluded. This happens due to the vessel becoming so thin that the segmentation breaks (thinner vessels are disrupted to a greater degree by smaller stenoses). Since both methods derive their measures from the binary segmentation mask all such cases are classed as occluded. This effect is seen most clearly on the 6 point scale. The GroBa method also shows a wider spread of values at lower grades, and mis-classifies a number of healthy vessels as high grades stenoses. This is likely due to premature triggering of the stopping criterion (the 2:1 aspect ratio of the estimated cylindrical balloon), leading to an underestimation of the vessel calibre as explored in Chapter 5.3. The CNN method is able to perform well over the entire severity range, achieving 74–92% accuracy on stenoses below 50% and 62–68% accuracy on grades above 50%. In nearly all cases, the CNN classifies correctly to within $\pm 1$ grade of the true value.

Reducing the fidelity of our scale to 3 points, as we recommended for our real ground truth data, the statistics improve, with a 57–96% accuracy for the Groba method, 81–94% for the cross-section method, and 86–94% for the CNN method. Similarly, reducing to a binary scale

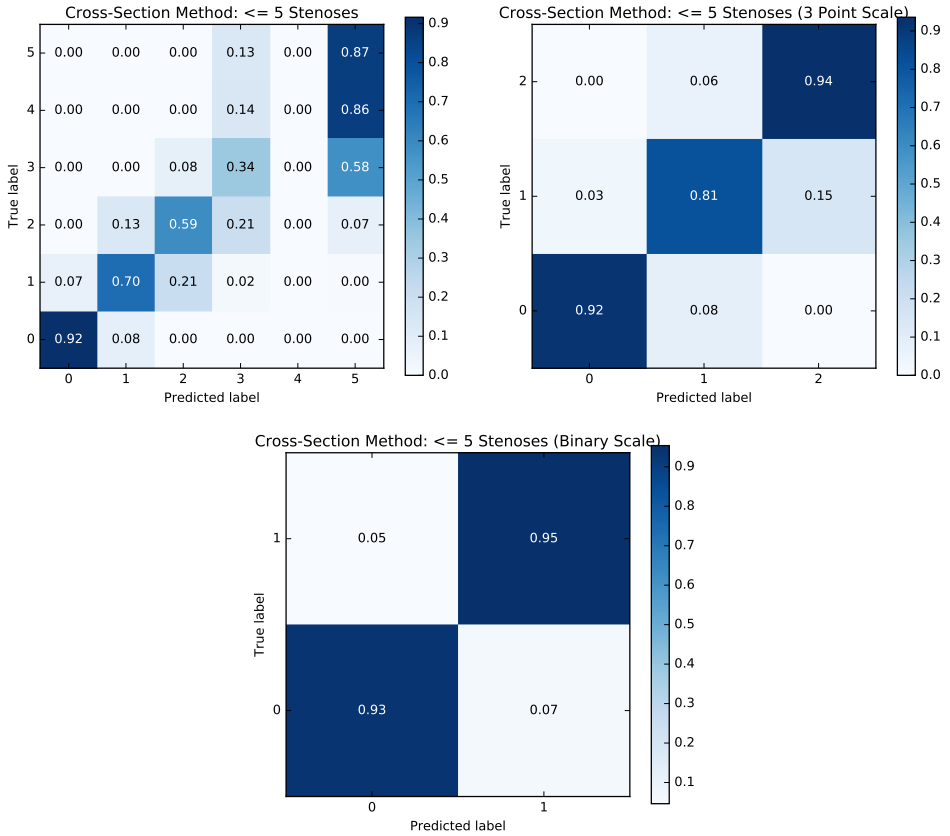



**Figure 8.5** – Confusion matrices for synthetic experiments with multiple stenoses ($\leq 5$), calculated with the cross-section method.

thresholded at 50% (the point at which stenoses are considered to be haemodynamically relevant) improve to 85–95% for GroBa, 93–95% for cross-sectional area, and 93–97% for the neural network, which are similar to the inter-rater agreement of the three annotators in Chapter 7.

In summary, these results show that when applied to synthetic vessels created to visually approximate the pre-processed subvolumes created by our pipeline, all three techniques perform to the same level as the clinicians on real data when examined on a binary scale. On the 3 point scale, the CNN and cross-section methods perform comparably, however the GroBa method has poor performance on the middle severity level, achieving an accuracy of only 57%. On the 6 point scale, the CNN method performs well over the entire range, achieving scores comparable to those of the agreement between annotator pairs on real data. We can therefore conclude that in principle, these methods (depending on the fidelity of the desired grading scheme) should be able to perform similarly to the real annotators used in our ground truth study provided that the complexity of the real data is similar to that of our synthetic set, the assumptions made in creating the synthetic dataset hold true (accurately tracked between junctions, normalised intensities, similar levels of vessel

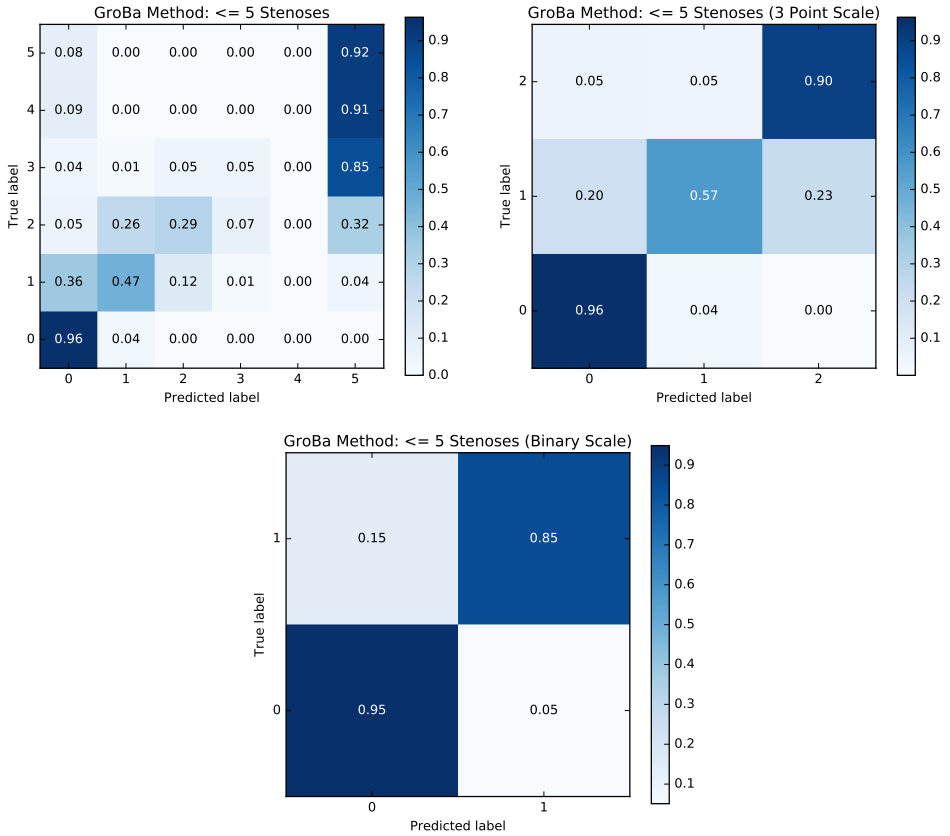



**Figure 8.6** – Confusion matrices for synthetic experiments with multiple stenoses ($\leq 5$), calculated with the GroBa method.

contrast, etc.), and in the case of the CNN that there is enough data to adequately train the network and capture the full range of variability seen.

## 8.4 Stenosis Grading with Real Data

In this section we will describe the results of applying the three stenosis grading algorithms, validated with synthetic data, to real vessels for which we have ground truth annotations. We will first describe the additional processing that was carried out in order to reduce artefacts and the variability of our set, before examining the final results and discussing the differences in performance when compared to our synthetic experiments.

### 8.4.1 Data Processing

Three additional aspects of the real data had to be addressed before applying our stenosis grading algorithms; locating the vessels to be extracted, removing extraneous material which is not part of the vessel of interest, and addressing the variable length of vessels for the CNN

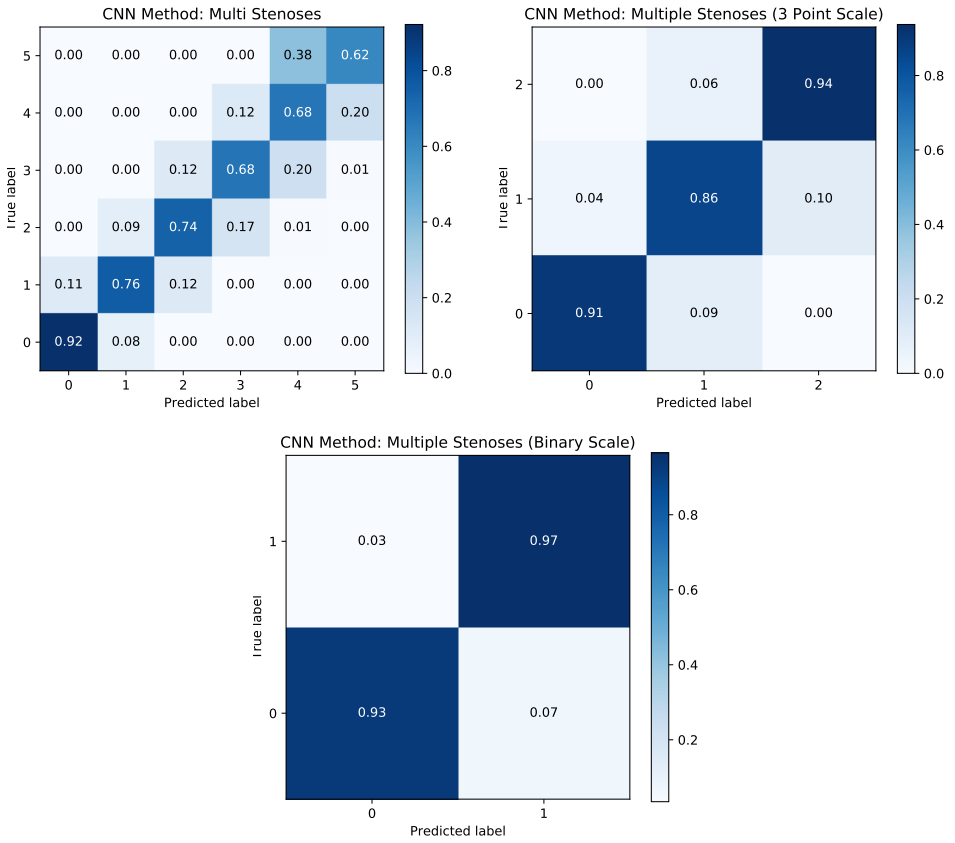



**Figure 8.7** – Confusion matrices for synthetic experiments with multiple stenoses ($\leq 5$), calculated with the convolutional neural network (CNN) method.

method which requires a constant input size.

### Locating Vessels of Interest

Automatically locating vessels with high accuracy is a complex task, with most published techniques focussing on detecting vessel landmarks [137, 138], though this has been shown to have high inter-rater variability for manual observers in WBMRA making ground truth validation difficult [130]. We instead adopt an automated procedure based on leveraging the annotation markers gathered during the ground truth collection exercise. We adopt the following steps for extracting a given vessel of interest.

1. All annotation markers (healthy, abnormal, and stenosis) from all three annotators are loaded for the given vessel using the vessel label assigned by the annotators.

2. Outliers are removed based on their coordinate positions in order to correct any wrongly assigned labels by the annotators.

3. A bounding box is calculated which encapsulates all the marker

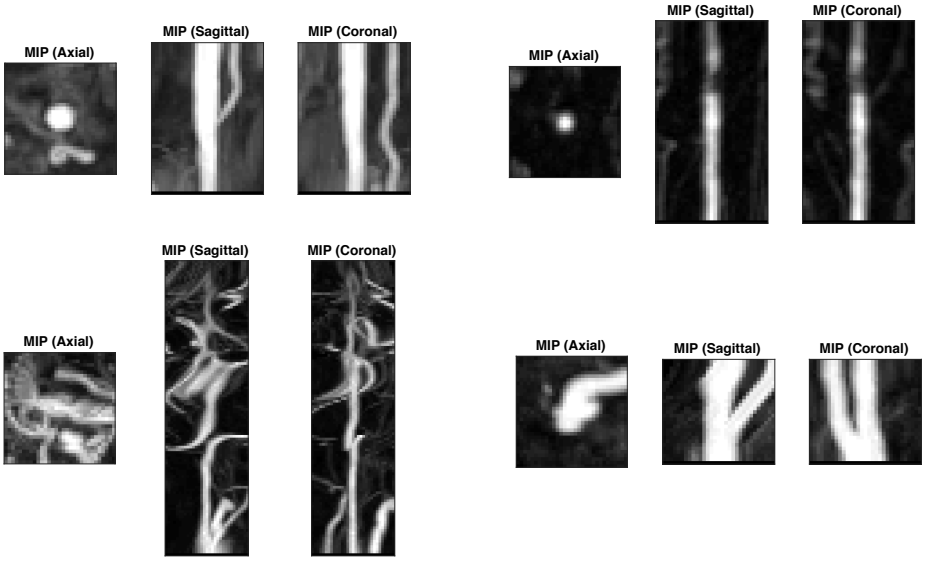



**Figure 8.8** – Four typical examples of a localised vessel volume processed with our pipeline, shown in the three standard planes.

coordinates, plus a border region of 5 voxels in case one of the border markers denotes a localised stenosis.

4. The vessel subvolume is extracted from the patient data volume and processed by our pipeline. This creates a segmentation mask (using our chosen method of OOF + threshold as described in Chapter 6) and corresponding centrelines for all vessels present in the subvolume.

5. Each annotation marker is linked to its closest centreline point.

6. The longest centreline with a linked marker point is selected as corresponding to the vessel of interest.

7. The vessel centreline is then tracked in 3D, and a stack of axial cross-sections used to create the vessel subvolume.

### Removing Extranous Material

Figure 8.8 shows four example vessels which have been processed by our pipeline. It can be seen that there are significant differences when compared to the idealised vessels of our synthetic dataset. In many cases there are additional vessel branches, nearby vessels or high intensity tissues outwith the tracked vessel of interest, and in some cases a significant "wobble" to the main vessel due to poor centreline tracking.

An additional processing step was therefore carried out in order to reduce the impact of these artefacts and allow our algorithms to be applied

without further modification. The approach taken was to extract only the central vessel from the volume, which we refer to as a "coring" procedure, i.e. the "core" of the vessel is retained while all other extraneous information is removed. This process consisted of three steps,

1. For each axial slice, a connected-component analysis is carried out on the corresponding segmentation mask, and only the component closest to the central voxel (i.e. the centreline point) is retained. This is based on the assumption that the centreline has tracked the vessel such that it is either inside or very close to the vessel of interest.

2. The mean diameter of the vessel is estimated by applying the cross-sectional area algorithm described in Section 8.2.1, taking the mean of the areas.

3. Assuming a circular cross-section of the vessel, a circular mask is created with an area equal to the previously calculated mean, and on each slice all voxels outside that mask are set to zero on both the raw data subvolume (for the CNN method), and the corresponding segmentation volume (for the cross-section and GroBa methods).

Figure 8.9 shows a comparison of pre- and post-processed vessel subvolumes using this technique. We can see that our coring technique is effective in removing extraneous information while retaining the vessel of interest.

### Variable Vessel Lengths

The GroBa and cross-sectional area algorithms can process vessels of arbitrary lengths, however the CNN algorithm expects input volumes to be of a constant dimension. The vessels which were included in our study range from a few tens of voxels in length (such as the renal arteries) to a few hundred voxels (such as the aorta and femoral arteries). We therefore constructed our subvolumes to have a cross-section large enough to encompass the aorta ($17 \times 17$). Splitting the vessels into equally sized parts without resampling could not be done as the ground truth label for each segment would no longer be know, therefore resampling was necessary. To achieve this a linear resampling of the data along the axial direction was performed, resizing each volume to a length equal to the mean vessel length across the entire set (68). This ensured that the volumes were all of equal size, and only involved resampling along a single axis.

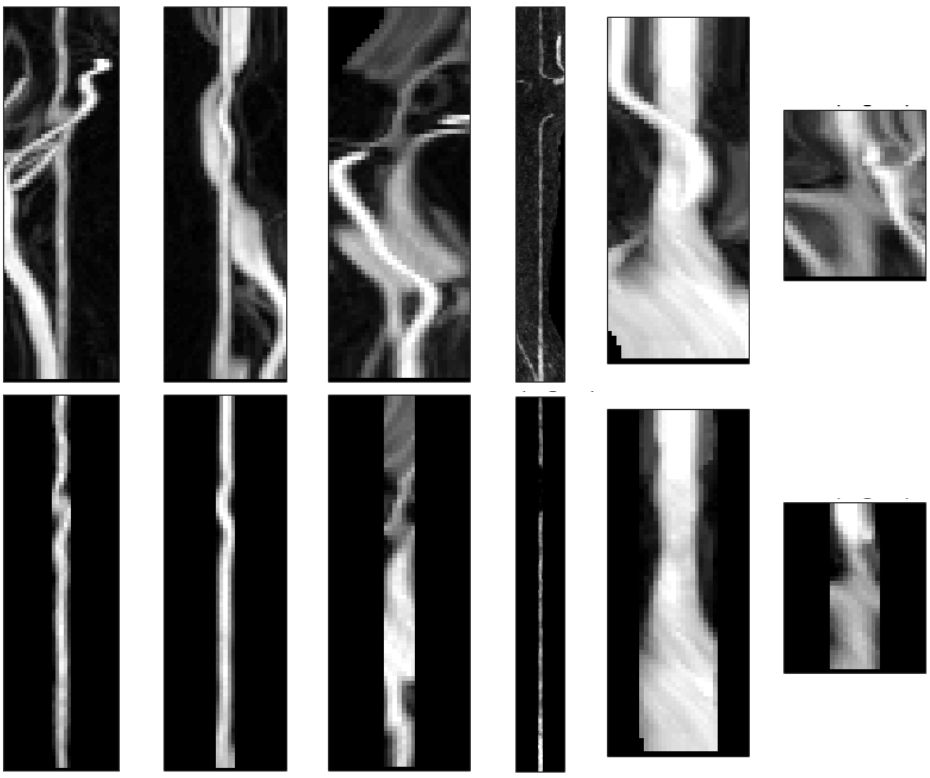


**Figure 8.9** – Typical examples of vessel volumes before (top) and after (bottom) processing by our "coring" technique to remove all information outside of the main vessel of interest.

### 8.4.2 Results

#### Vessel Tracking

We applied the data processing pipeline described in Section 8.4.1 to all 666 vessels for which we have ground truth annotations. A lower threshold of 10 voxels was set (around half the shortest vessel), below which the algorithm was deemed to have failed to detect a reliable vessel and no subvolume was extracted. In around 6% of cases the algorithm failed to extract a subvolume. Figure 8.10 shows a histogram of the vessels which failed, and the grades which were assigned. These were spread evenly across all patients, and across a range of vessels. A possible reason for this can be seen in the grade histogram, where the majority of failed extractions came from healthy vessels. For these cases very few annotation points existed, typically just a single point at the start and end of the vessel denoting a healthy section along its entire length. Since these were often placed very close to junctions, it is likely that in many of these cases a small centreline near the junction was picked which led to a failed extraction. Other potential reasons are poor vessel tracking

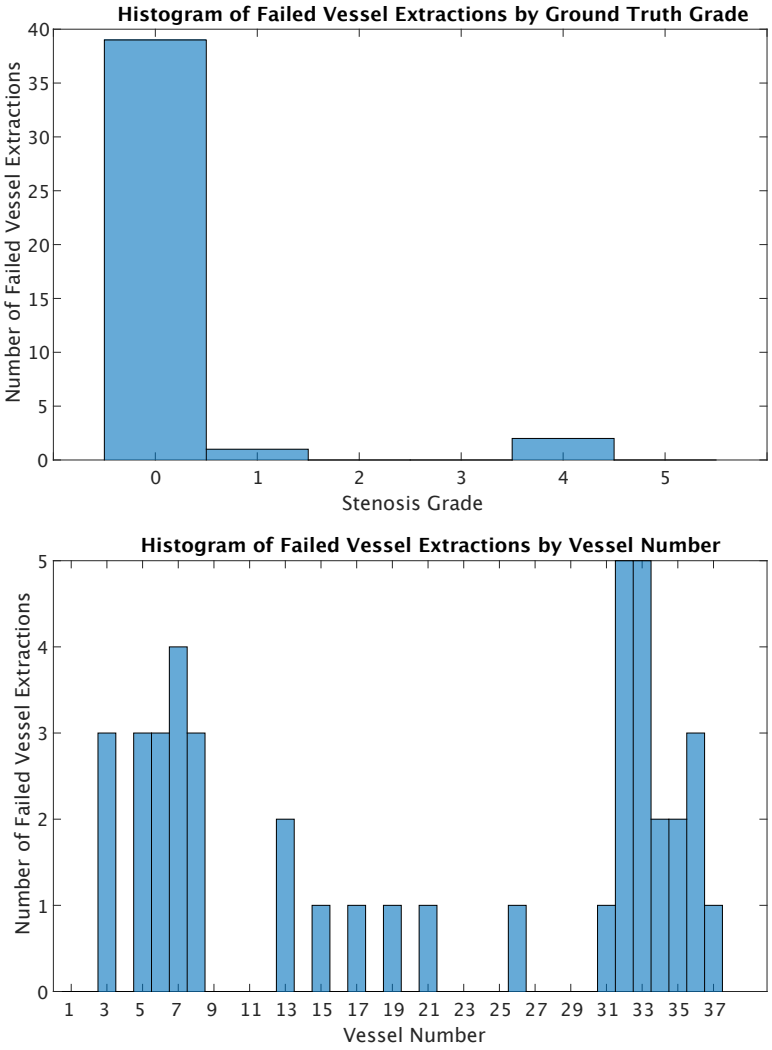



**Figure 8.10** – Histograms of the vessels which were unable to be extracted using our method, separated by grade (top) and vessel label (bottom).

leading to short, disjointed sections, poor vessel contrast in certain scans (particularly towards the extremities) leading to inaccurate segmentations, and a lack of annotations for a particular vessel (sometimes just one or none at all) due to them being deemed of non-diagnostic quality by the annotators.

### Stenosis Grading

All successfully processed vessels were analysed by the cross-section, GroBa, and CNN stenosis grading algorithms. It was found that the CNN algorithm could not be trained adequately from scratch using our small number of extracted vessels. Shallower networks led to unstable training performance which failed to converge, and varied markedly depending on the random initialisation of weights. Deeper networks quickly fell into sub-optimal solutions, classifying all subvolumes as a severity of

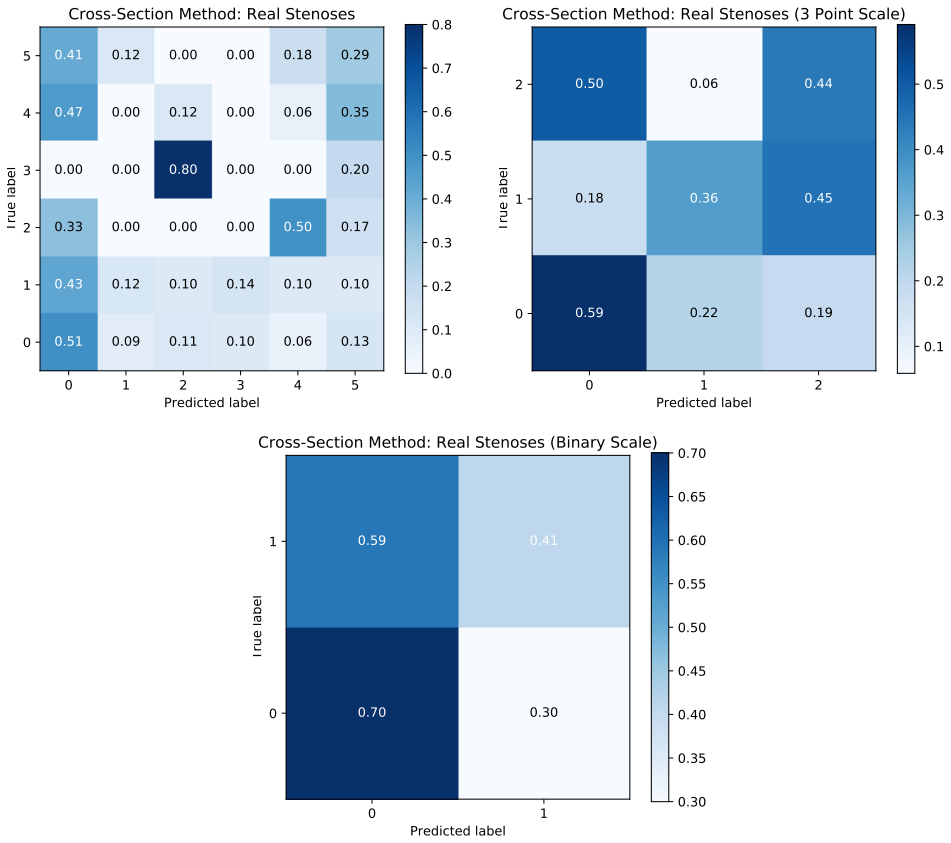



**Figure 8.11** – Confusion matrices for real vessel data, calculated with the cross-section method.

zero.

In order to ameliorate these issues linked to training from scratch with limited data, we adopt a pre-training step using the previously generated synthetic vessels. Pre-trained networks are used in many areas of deep learning research, particularly where very deep networks such as VGGNet [139] and GoogLeNet [140] can be trained on databases of millions of natural images before being refined to a specific task through transfer learning [141,142]. No such networks currently exist for 3D medical imaging however, and so an alternative method is needed. Pre-training with synthetic data has been used in a variety of applications, such as 3D scene understanding [143], text localisation in natural images [144], object detection [145], 3D face reconstruction [146], and segmentation inpainting for plant root networks [147], among others. In a similar manner, we opted to use the weights of the network trained on our synthetic vessel dataset as our pre-trained network, allowing it to be trained to a stable (albeit poor) solution with our limited real data. The re-training was done in 6 folds, with the vessels from 3 patients being held for testing, and the training set in each fold being balanced by data augmentation through rotation and mirroring and 5% used for validation.

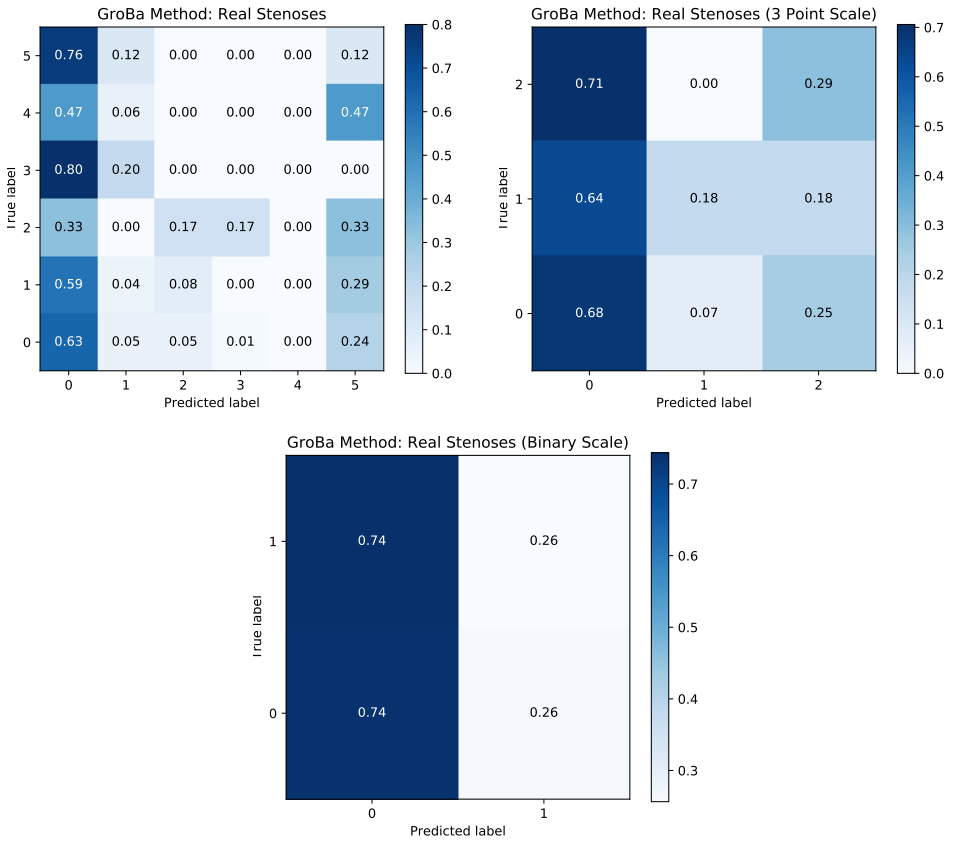



**Figure 8.12** – Confusion matrices for real vessel data, calculated with the GroBa method.

The results of applying the cross-section, GroBa, and CNN stenosis grading algorithms are shown in Figures 8.11, 8.12, and 8.13 respectively. We can see that the performance of all three approaches is poor as compared to our synthetic experiments. There is a tendency of all three algorithms to classify most vessels as healthy, even those of high stenosis grade. There are a number of possible reasons for this.

- The tracking approach we adopted is derived from the segmentation mask, and high grade stenoses have a tendency to disrupt the segmentation. Therefore vessels which contain a severe stenosis may only be tracked up to the point of stenosis and not through it, giving the appearance of a healthy vessel.

- We have a limited number of vessel volumes for training the CNN method, the majority of which are healthy (around 3/4), so our set does not capture the full range of vessel and stenosis morphologies.

- The ground truth labels themselves also have inherent noise, as described in Chapter 7.

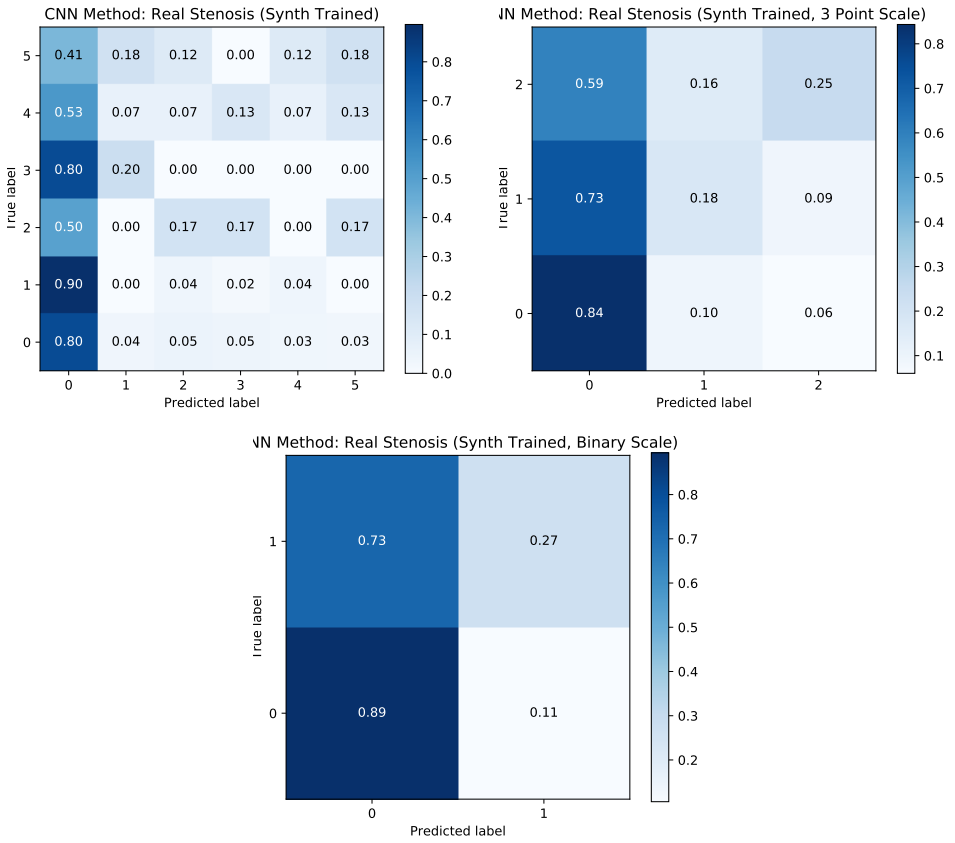



**Figure 8.13** – Confusion matrices for real vessel data, calculated with the convolutional neural network (CNN) method.

## 8.5 Conclusions

Here we have presented three different algorithms for grading stenosis severity of vessel segments in WBMRA. Two are based on a set of rules applied to a set of calibre measurements — cross-sectional area and GroBa calibre — with the third using a convolutional neural network to grade entire vessel segments based on the maximum stenosis present.

A set of ten thousand synthetic vessels were created, with a set of randomly varied parameters which were chosen to give a close visual resemblance to real vessels in our dataset. A number of simplifying assumptions had to be made, such as no branches or non-stenotic abnormalities. Applying our algorithms to this synthetic set showed good performance, achieving similar performance to the inter-rater agreement between annotators on real data when examined at the binary level for GroBa, the 3-point scale for the cross-section method, and the original 6-point scale for the CNN.

Applying the stenosis algorithms to real data showed much poorer performance. All three had a tendency to over-classify vessels as healthy, with false negative rates of 59%, 74%, and 73% for the cross-section,

GroBa, and CNN methods respectively.

We conclude that while the methods performed well on synthetic vessels, the real data produced by our processing pipeline is significantly more complex. We believe the main limiting factors which contributed to this was the difficulty tracking vessels in the presence of severe disease and the paucity (and variability) of ground truth data available. We will address each of these issues, and how we recommend improving upon them in the future, in Chapter 9.4.

# Chapter 9

# Conclusions and Future Work

## 9.1 Introduction

In this chapter we will summarise the work presented in this thesis, discussing our main contributions along with the limitations of our work. We will conclude with a discussion of how this work could be extended in the future, and provide recommendations based on the results we have presented.

## 9.2 Summary of Thesis

In this thesis we have proposed an automated framework for analysing vessels in WBMRA examinations, and assigning them a grade based on the maximum severity stenosis detected. This involved constructing a processing pipeline which would take a pre- and post-contrast MRA volume as it's input, automatically segment the arteries, calculate their centrelines to track them in 3D space, then detect stenoses within the tracked vessel volumes.

To accomplish this, we first began with an interface prototyping exercise in collaboration with our clinical and industry partners, detailed in Chapter 3. This allowed us to identify the key demands of the clinical task, and the technologies needed to meet them. Here it was concluded that the key factor from clinicians was that the tool should be be fully automatic and require no manual intervention, and should be targeted as a tool for flagging up areas of suspected disease so as to direct clinicians attention during the diagnosis process. This targeted our research towards two key areas; a study of 3D vessel segmentation algorithms validated on ground truth WBMRA vessel maps, and development of automated stenosis detection algorithms validated against ground truth

stenosis severity markings.

For this project, 18 whole-body patient datasets were provided, acquired at Dundee's Ninewells Hospital as part of a separate study [83], and were selected to give a range of healthy volunteers and patients with mild to severe peripheral arterial disease. The first task undertaken was to compare two pre-packaged software solutions for volume registration of pre- and post-contrast datasets, described in Chapter 5.2. We chose to use the "tmvsFire" software developed by our industrial partner, which gave the fewest artefacts after digital subtraction of the two datasets subtraction. Setting this pre-processing step at the start ensured that all future work would be carried out on the same set of subtracted data volumes.

Next we evaluated a baseline method for vessel analysis in WBMRA called the "GroBa" system, published in 2012. The performance of this system had not previously been fully characterised, so a number of synthetic test were carried out, alongside a comparison with a number of manual stenosis measurements, as described in Chapter 5.3. It was found to perform well for simple models, but a number of issues were highlighted when examining more complex and real-world examples. The main limitations identified was the poor segmentation performance (using the Frangi filter), and issues with early triggering of the stopping criteria resulting in a wrong estimation of vessel calibre.

In Chapter 6 we addressed the segmentation shortcomings of the baseline system by performing a quantitative comparison between three previously published methods (Frangi, active contours, and OOF), and our developed method using a convolutional neural network. We created a detailed segmentation protocol which was followed to create three ground truth vessel maps by a trained annotator with previous experience annotating medical images. Our initial experiments revealed that the OOF algorithm provided the best mean Dice score of 0.705 across all stations for our data, narrowly beating the CNN method due to the small amount of training data available. Due to the chronology of the work carried out in this project, the OOF method was used for all future work. However, subsequent experiments revealed that a 2D U-Net architecture outperformed the OOF method, achieving a mean Dice score of 0.756 as described in Chapter 6.6.

A further ground truth data collection task was then undertaken in Chapter 7 to manually annotate the locations of all stenoses within our 18 patient datasets. We constructed a clear annotation protocol in conjunc-

tion with three trained clinicians, using research software provided by the Canon Medical team. When comparing results between the three annotators we found significant inter-observer variability even at the highest severity levels. A fourth experienced clinical expert was asked to review the 37 highest disagreement cases, providing a definitive assessment of the true score alongside the likely reasoning for the disagreements. These included user Error, image artefacts, poor image quality, natural anatomical variability, and simple differences of opinion. Based on these results we re-analysed the full dataset at three different levels of fidelity — a 6 point, 3-point, and binary scale — and found that the annotation approach that was taken (assessment by eye) was most suited to a 3-point severity scale, where inter-rater agreement improved from 72.5% to 92.5%.

In Chapter 8 we present three algorithms for detecting stenoses within tracked vessels; two rule-based methods using cross-sectional areas and GroBa calibres, and one machine learning based method in the form of an 8 layer CNN. We first constructed a synthetic experiment, generating a dataset of ten thousand vessel volumes against which the algorithms were applied. The parameters for these vessels were selected to give a similar appearance to the real processed vessels given some simplifying constraints (such as no junctions, and no abnormalities that aren't stenoses). All three algorithms gave good results when examined in the same way as the ground truth annotations, achieving similar performance to the inter-rater agreement between annotators on real data when examined at the binary level for GroBa, the 3-point scale for the cross-section method, and the original 6-point scale for the CNN.

Applying our algorithms to the vessel volumes extracted from our 18 patients using the ground truth stenosis annotations we collected we get poor performance. All methods tend to over-classify vessels as healthy even in the binary case, with true negative and false negative rates of 70%/59%, 74%/74% and 89%/73% for the cross-section, GroBa, and CNN methods respectively. The main issues which must be addressed in future to improve this performance is the vessel segmentation and tracking around areas of high stenosis, and increasing the amount of ground truth data available for training.

## 9.3 Contributions

In this thesis we put forward a fully automated framework for the processing and analysis of WBMRA data for grading stenoses. We contribute to the existing literature of automated analysis of WBMRA through the quantitative comparison of vessel segmentation techniques, validated against manual ground truth, and through quantifying the inter- and intra-observer variability of manual stenosis grading by trained observers, providing a clear protocol and recommended guidelines for collecting annotations to be used as ground truth data for algorithm development. The key contributions of this thesis can be summarised as follows,

1. First quantitative comparison of vessel segmentation techniques for WBMRA data, including the first application of a 3D convolutional neural network to the task of automatic vessel segmentation in WBMRA.

2. Quantification of the inter-observer variability for the task of manually detecting and grading stenoses from WBMRA examinations by trained clinicians.

3. Development of clear protocols for the collection of ground truth data for vessel segmentation and stenosis severity in WBMRA volumes, putting forward good-practice guidelines for ground truth collection from WBMRA data for algorithm development.

4. First fully automated framework for extracting and analysing arteries in WBMRA examinations to be tested against well-characterised ground truth stenosis annotations in real patient data.

## 9.4 Limitations and Future Work

We will conclude by discussing the main limitations of our proposed system, and possible solutions and extensions that could be explored in the future. As discussed in Chapter 3.3, the target of assistive software tools for CVD diagnosis in WBMRA data should focus on a fully automated system which reduces the clinical workload by directing the attention of clinicians to the areas of the vasculature which are detected to be abnormal. In order for the system proposed in this thesis to meet that goal, we have identified a number of areas which should be the focus of future research.

### 9.4.1 Vessel Segmentation

One of the main difficulties encountered when attempting to grade vessels in our dataset was the issue of poor tracking around areas of high stenosis. Our segmentation comparison was only validated against ground truth vessel maps for three patients, due to the time investment required to construct the protocol and manually annotate the large volumes, slice by slice. The three patients chosen for annotation were also from the healthy and low severity groups in order to simplify the annotation task, however this also meant that pathology was not well captured in the data. We therefore recommend that further ground truth data collection should be targeted, acquired following the optimised protocol we have presented. This would be especially useful for the application of deeper neural networks.

Following our initial results from the U-Net network, we also recognise that a deeper comparison of network architectures could improve the segmentation performance further. One approach would be to include the output of the OOF algorithm as an additional input into a the network, either as a single channel with the maximum OOF response or multiple channels containing the OOF response at each radius step. A thorough comparison of 2D U-Net and 3D V-Net architectures alongside more complex data augmentation techniques may also yield improved performance, though these would still be subject to the same constraints imposed by the lack of training data, making further ground truth segmentation collection the logical first target for future work.

### 9.4.2 Vessel Tracking

In this project we opted to use an off-the-shelf solution for automatic vessel tracking which calculates vessel centrelines from binary segmentation masks. It was seen in examples from our data that this often performed poorly in some situations due to its reliance on the segmentation. In particular, stenoses tend to disrupt or break the normal path of the vessels leading to inaccurate tracking through areas of severe disease. A more sophisticated method will need to be developed to give robust tracking in the presence of pathology.

Tracking across severe stenoses or occlusions also poses an issue, since other contextual information must be leveraged in order to accurately "reconnect" the centrelines across the broken vessels. A landmark-based tracking system, following such work as presented in [137,148] may be one

possible solution, or an atlas-based approach such as explored in [138]. However both approaches would need significant amounts of ground truth in order to be robust to both the range of natural anatomical variations and more importantly, the variability of pathology.

### 9.4.3 Ground Truth Stenosis Data

The ground truth collection exercise we carried out with clinicians for annotating stenoses in WBMRA raised a number of unforeseen issues which had not previously been reported or explored in the literature, including 37 cases (almost 6% of examined vessels) which contained a disagreement of at least 2 grades involving a stenosis considered to be haemodynamically important ($> 50\%$).

Further ground truth data is clearly needed to fully characterise the range of pathology seen across the population, and provide more substantially more data for algorithm development. We recommend that this should be a key target for future work, with annotation tasks carried out following our proposed guidelines in Chapter 7, with all vessel grades assigned following the consensus of multiple trained annotators on a 3 point grading scale (healthy, low severity, high severity).

### 9.4.4 Automated Stenosis Grading

The three algorithms proposed in Chapter 8 gave promising results in synthetic experiments, showing that they should work in principle if the complexity of our synthetic data was comparable to that of the vessels extracted from our 18 patients. When applied to the real data the final performance of all three approaches was poor however. Through examining the appearance of the extracted vessels, and performance of our processing pipeline in the presence of complex pathology, we deduce that the key limiting factors to this study was the accuracy of vessel segmentation and tracking (particularly in the presence of disease), and the quantity and quality of ground truth data for training. Our proposals for future developments in Sections 9.4.1, 9.4.2, and 9.4.3 above should help improve these factors greatly. A greater pool of training data as proposed in Section 9.4.3 would also allow the training of a more complex CNN algorithm, and enable a more robust assessment of our proposed stenosis grading algorithms a wider number of vessel examples.

# Bibliography


[1] D. W. McRobbie, E. A. Moore, M. J. Graves, and M. R. Prince, *MRI from Picture to Proton*. Cambridge University Press, second edition ed., 2007.

[2] H. Kramer, H. Quick, B. Tombach, S. Schoenberg, and J. Barkhausen, "Whole-body MRA," *European Radiology*, vol. 18, no. 9, pp. 1925–1936, 2008.

[3] J. Weir-McCall, S. Duce, S. Gandy, S. Matthew, P. Martin, D. Cassidy, L. McCormick, J. Belch, A. Struthers, H. Colhoun, and J. Houston, "Whole body cardiovascular magnetic resonance imaging to stratify symptomatic and asymptomatic atherosclerotic burden in patients with isolated cardiovascular disease," *BMC Medical Imaging*, vol. 16, 2016.

[4] R. Van Uitert and I. Bitter, "Subvoxel precise skeletons of volumetric data based on fast marching methods," *Medical Physics*, vol. 34, no. 2, pp. 627–638, 2007.

[5] S.-R. Wang, C.-L. Tung, M.-T. Wu, and Y.-N. Sun, "Dual snake-based vessel wall detection for MSCT coronary artery analysis," *Journal of Medical and Biological Engineering*, vol. 27, no. 4, pp. 165–171, 2007.

[6] A. Suinesiaputra, P. Koning, E. Zudilova-Seinstra, J. Reiber, and R. Geest, "Automated quantification of carotid artery stenosis on contrast-enhanced MRA data using a deformable vascular tube model," *The International Journal of Cardiovascular Imaging*, vol. 28, no. 6, pp. 1513–1524, 2012.

[7] N. Zhu and A. Chung, "Graph-based optimal cross section boundary for vessel segmentation and stenosis quantification," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pp. 2041–2044, Sep 2012.

[8] J. A. Ambrose, S. L. Winters, A. Stern, A. Eng, L. E. Teicholz, R. Gorlin, and V. Fuster, "Angiographie morphology and the pathogenesis of unstable angina pectoris," *Journal of the American College of Cardiology*, vol. 5, no. 3, pp. 609–616, 1985.

[9] A. Perez-Rovira, E. Trucco, J. Weir-McCall, and G. Houston, "GroBa: Growing balloons for calibre measurement on stenotic lumens," in *Computer-Based Medical Systems (CBMS), 2012 25th International Symposium on*, pp. 1–6, 2012.

[10] E. Wilkins, L. Wilson, K. Wickramasinghe, P. Bhatnagar, J. Leal, R. Luengo-Fernandez, R. Burns, M. Rayner, and N. Townsend, *European Cardiovascular Disease Statistics 2017 Edition*. European Heart Network, Brussels, 2017.

[11] M. de Jong, H. Worp, Y. Graaf, F. Visseren, and J. Westerink, "Pioglitazone and the secondary prevention of cardiovascular disease. A meta-analysis of randomized-controlled trials," *Cardiovascular Diabetology*, vol. 16, 12 2017.

[12] C. Lundberg, L. Johansson, C. E. Barbier, L. Lind, H. Ahlström, and T. Hansen, "Total atherosclerotic burden by whole body magnetic resonance angiography predicts major adverse cardiovascular events," *Atherosclerosis*, vol. 228, no. 12, pp. 148–152, 2013.

[13] S. G. Ruehm, S. C. Goehde, and M. Goyen, "Whole body MR angiography screening," *The International Journal of Cardiovascular Imaging*, vol. 20, no. 6, pp. 587–591, 2004.

[14] Z. Cai, A. Stolpen, M. Sharafuddin, R. McCabe, H. Bai, T. Potts, M. Vannier, D. Li, X. Bi, J. Bennett, J. Golzarian, S. Sun, G. Wang, and E.-W. Bai, "Bolus characteristics based on magnetic resonance angiography," *Biomedical engineering online*, vol. 5, p. 53, 2006.

[15] Withinsight Framework by Claron Technology Inc., http://www.clarontech.com.

[16] COR Analyzer System by Rcadia Medical Imaging Ltd., http://www.rcadia.com.

[17] VesselMASS RE by Medis Specials, http://www.medisspecials.com.

[18] K. Anders, S. Achenbach, I. Petit, W. Daniel, M. Uder, and T. Pflederer, "Accuracy of automated software-guided detection of significant coronary artery stenosis by CT angiography: comparison with invasive catheterisation," *European Radiology*, vol. 23, no. 5, pp. 1218–1225, 2013.

[19] I. M. Adame, P. J. de Koning, B. P. Lelieveldt, B. A. Wasserman, J. H. Reiber, and R. J. van der Geest, "An integrated automated analysis method for quantifying vessel stenosis and plaque burden from carotid MRI images: Combined postprocessing of MRA and vessel wall MR," *Stroke*, vol. 37, no. 8, pp. 2162–2164, 2006.

[20] C. Kirbas and F. Quek, "A review of vessel extraction techniques and algorithms," *ACM Comput. Surv.*, vol. 36, pp. 81–121, Jun 2004.

[21] D. Lesage, E. D. Angelini, I. Bloch, and G. Funka-Lea, "A review of 3D vessel lumen segmentation techniques: Models, features and extraction schemes," *Medical Image Analysis*, vol. 13, no. 6, pp. 819–845, 2009.

[22] S. M. Debrey, H. Yu, J. K. Lynch, K.-O. Lövblad, V. L. Wright, S.-J. D. Janket, and A. E. Baird, "Diagnostic accuracy of magnetic resonance angiography for internal carotid artery disease: A systematic review and meta-analysis," *Stroke*, vol. 39, no. 8, pp. 2237–2248, 2008.

[23] M. E. Westwood, S. Kelly, E. Berry, J. M. Bamford, M. J. Gough, C. M. Airey, J. F. M. Meaney, L. M. Davies, J. Cullingworth, and M. A. Smith, "Use of magnetic resonance angiography to select candidates with recently symptomatic carotid stenosis for surgery: systematic review," *BMJ*, vol. 324, p. 198, Jan 2002.

[24] G. Blackwell, S. Wann, and S. Kadekar, "Clinical applications of cardiovascular magnetic resonance methods," *Journal of Tehran University Heart Center*, vol. 1, no. 3, 2011.

[25] G. Schmidt, D. Dinter, M. F. Reiser, and S. O. Schoenberg, "The uses and limitations of whole-body magnetic resonance imaging," *Dtsch Arztebl International*, vol. 107, no. 22, pp. 383–389, 2010.

[26] F. Bamberg, K. G. Parhofer, E. Lochner, R. P. Marcus, D. Theisen, H. M. Findeisen, U. Hoffmann, S. O. Schönberg, C. L. Schlett,

M. F. Reiser, and S. Weckbach, "Diabetes mellitus: Long-term prognostic value of whole-body MR imaging for the occurrence of cardiac and cerebrovascular events," *Radiology*, vol. 269, no. 3, pp. 730–737, 2013.

[27] P. Reimer, P. M. Parizel, J. F. M. Meaney, and F. A. Stichnoth, *Clinical MR Imaging: A Practical Approach*. Springer Berlin Heidelberg, third edition ed., 2010.

[28] "Cardiovascular disease." https://www.nhs.uk/conditions/cardiovascular-disease/. Accessed: 2019-10-30.

[29] D. C. Goff, D. M. Lloyd-Jones, G. Bennett, S. Coady, R. B. D'Agostino, R. Gibbons, P. Greenland, D. T. Lackland, D. Levy, C. J. O'Donnell, J. G. Robinson, J. S. Schwartz, S. T. Shero, S. C. Smith, P. Sorlie, N. J. Stone, and P. W. Wilson, "2013 acc/aha guideline on the assessment of cardiovascular risk," *Journal of the American College of Cardiology*, vol. 63, no. 25, pp. 2935–2959, 2014.

[30] M. Al-Qaisi, D. Nott, D. King, and S. Kaddoura, "Ankle brachial pressure index (ABPI): An update for practitioners," *Vasc Health Risk Manag.*, vol. 5, pp. 833–841, 2009.

[31] N. S. Chahal, T. K. Lim, P. Jain, J. C. Chambers, J. S. Kooner, and R. Senior, "Does subclinical atherosclerosis burden identify the increased risk of cardiovascular disease mortality among united kingdom indian asians? A population study," *American Heart Journal*, vol. 162, no. 3, pp. 460–466, 2011.

[32] S. Gandy, M. Lambert, J. Belch, I. Cavin, E. Crowe, R. Littleford, J. Macfarlane, S. Matthew, P. Martin, R. Nicholas, A. Struthers, F. Sullivan, S. Waugh, R. White, J. Weir-McCall, and J. Houston, "Technical assessment of whole body angiography and cardiac function within a single MRI examination," *Clinical Radiology*, vol. 70, no. 6, pp. 595–603, 2015.

[33] S. L. Duce, J. R. Weir-McCall, S. J. Gandy, S. Z. Matthew, D. B. Cassidy, L. McCormick, P. Rauchhaus, H. Looker, H. M. Colhoun, and J. G. Houston, "Cohort comparison study of cardiac disease and atherosclerotic burden in type 2 diabetic adults using whole body cardiovascular magnetic resonance imaging," *Cardiovasc Diabetol*, vol. 14, no. 22, 2015.

[34] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy, "Supervised learning from multiple experts: Whom to trust when everyone lies a bit," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 889–896, ACM, 2009.

[35] Y. Yoo, T. Brosch, A. Traboulsee, D. Li, and R. Tam, "Deep learning of image features from unlabeled data for multiple sclerosis lesion segmentation," in *Machine Learning in Medical Imaging* (G. Wu, D. Zhang, and L. Zhou, eds.), vol. 8679 of *Lecture Notes in Computer Science*, pp. 117–124, 2014.

[36] L. Joskowicz, D. Cohen, N. Caplan, and J. Sosna, "Inter-observer variability of manual contour delineation of structures in CT," *European Radiology*, Sep 2018.

[37] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab, "AggNet: Deep learning from crowds for mitosis detection in breast cancer histology images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1313–1321, 2016.

[38] A. Q. O'Neil, J. T. Murchison, E. J. R. van Beek, and K. A. Goatman, "Crowdsourcing labels for pathological patterns in CT lung scans: Can non-experts contribute expert-quality ground truth?," in *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pp. 96–105, 2017.

[39] R. Kwitt, S. Hegenbart, N. Rasiwasia, A. Vécsei, and A. Uhl, "Do we need annotation experts? A case study in celiac disease classification," in *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2014*, vol. 8674 of *Lecture Notes in Computer Science*, pp. 454–461, Springer International Publishing, 2014.

[40] D. Mitry, K. Zutis, B. Dhillon, T. Peto, S. Hayat, K.-T. Khaw, J. E. Morgan, W. Moncur, E. Trucco, and P. J. Foster, "The accuracy and reliability of crowdsource annotations of digital retinal images," *Translational Vision Science and Technology*, vol. 5, no. 5, p. 6, 2016.

[41] A. Kovashka, O. Russakovsky, and L. Fei-Fei, *Crowdsourcing in Computer Vision*. Hanover, MA, USA: Now Publishers Inc., 2016.

[42] M. Guillaumin, D. Kuttel, and V. Ferrari, "ImageNet auto-annotation with segmentation propagation," *International Journal of Computer Vision*, vol. 110, no. 3, pp. 328–348, 2014.

[43] V. V. Valindria, I. Lavdas, W. Bai, K. Kamnitsas, E. O. Aboagye, A. G. Rockall, D. Rueckert, and B. Glocker, "Reverse classification accuracy: Predicting segmentation performance in the absence of ground truth," *IEEE Transactions on Medical Imaging*, vol. 36, no. 8, pp. 1597–1606, 2017.

[44] A. Frangi, W. Niessen, K. Vincken, and M. Viergever, "Multi-scale vessel enhancement filtering," in *Medical Image Computing and Computer-Assisted Interventation — MICCAI 1998* (W. Wells, A. Colchester, and S. Delp, eds.), vol. 1496 of *Lecture Notes in Computer Science*, pp. 130–137, Springer Berlin Heidelberg, 1998.

[45] Y. Sato, S. Nakajima, H. Atsumi, T. Koller, G. Gerig, S. Yoshida, and R. Kikinis, "3D multi-scale line filter for segmentation and visualization of curvilinear structures in medical images," in *CVRMed-MRCAS'97* (J. Troccaz, E. Grimson, and R. Mösges, eds.), vol. 1205 of *Lecture Notes in Computer Science*, pp. 213–222, Springer Berlin Heidelberg, 1997.

[46] H. E. Bennink, H. C. Van Assen, G. J. Streekstra, R. T. Wee, J. A. E. Spaan, and B. M. T. H. Romeny, "A novel 3D multi-scale lineness filter for vessel detection," in *Proceedings of the 10th International Conference on Medical Image Computing and Computer-assisted Intervention*, MICCAI'07, (Berlin, Heidelberg), pp. 436–443, Springer-Verlag, 2007.

[47] J. Hannink, R. Duits, and E. Bekkers, "Crossing-preserving multi-scale vesselness," in *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2014* (P. Golland, N. Hata, C. Barillot, J. Hornegger, and R. Howe, eds.), vol. 8674 of *Lecture Notes in Computer Science*, pp. 603–610, Springer International Publishing, 2014.

[48] M. Law and A. Chung, "Three dimensional curvilinear structure detection using optimally oriented flux," in *Computer Vision ECCV 2008*, vol. 5305 of *Lecture Notes in Computer Science*, pp. 368–382, Springer Berlin Heidelberg, 2008.

[49] E. Türetken, C. Becker, P. Glowacki, F. Benmansour, and P. Fua, "Detecting irregular curvilinear structures in gray scale and color imagery using multi-directional oriented flux," in *2013 IEEE International Conference on Computer Vision*, pp. 1553–1560, 2013.

[50] B. Abdollahi, A. El-Baz, and A. Amini, "A multi-scale non-linear vessel enhancement technique," in *Engineering in Medicine and Biology Society,EMBC, 2011 Annual International Conference of the IEEE*, pp. 3925–3929, 2011.

[51] J. Brieva, E. Gonzalez, F. Gonzalez, A. Bousse, and J. J. Bellanger, "A level set method for vessel segmentation in coronary angiography," in *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, pp. 6348–6351, Jan 2005.

[52] X. Gao, Y. Uchiyama, X. Zhou, T. Hara, T. Asano, and H. Fujita, "A fast and fully automatic method for cerebrovascular segmentation on time-of-flight (TOF) MRA image," *Journal of Digital Imaging*, vol. 24, no. 4, pp. 609–625, 2011.

[53] B. Mohr, S. Masood, and C. Plakas, "Accurate lumen segmentation and stenosis detection and quantification in coronary CTA," in *Proceedings of 3D Cardiovascular Imaging: a MICCAI Segmentation Challenge Workshop.*, 2012.

[54] N. D. Forkert, A. Schmidt-Richberg, J. Fiehler, T. Illies, D. Möller, D. Säring, H. Handels, and J. Ehrhardt, "3D cerebrovascular segmentation combining fuzzy vessel enhancement and level-sets with anisotropic energy weights," *Magnetic Resonance Imaging*, vol. 31, no. 2, pp. 262–271, 2013.

[55] J. A. Sethian, *Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science*, vol. 3. Cambridge University Press, 1999.

[56] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1097–1105, 2012.

[57] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[58] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III*, pp. 234–241, 2015.

[59] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 640–651, April 2017.

[60] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[61] A. A. Farag, M. S. Hassouna, R. Falk, and S. Hushek, "Reliable fly-throughs of vascular trees," tech. rep., Computer Vision and Image Processing Laboratory, Department of Electrical and Computer Engineering, University of Louisville, 2004.

[62] K. Krissian, R. Kikinis, and C.-F. Westin, "Algorithms for extracting vessel centerlines," tech. rep., Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Sep 2004.

[63] G. Borgefors, "Distance transformations in arbitrary dimensions," *Computer Vision, Graphics, and Image Processing*, vol. 27, pp. 321–345, 1984.

[64] M. M. Macedo, C. Mekkaoui, and M. P. Jackowski, "Vessel centerline tracking in CTA and MRA images using hough transform," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* (I. Bloch and J. Cesar, RobertoM., eds.), vol. 6419 of *Lecture Notes in Computer Science*, pp. 295–302, Springer Berlin Heidelberg, 2010.

[65] F. Zamani Boroujeni, R. Rahmat, N. Mustapha, L. Affendey, and O. Maskon, "Coronary artery center-line extraction using second order local features," *Computational and mathematical methods in medicine*, vol. 2012, p. 940981, 2012.

[66] S. Worz and K. Rohr, "Segmentation and quantification of human vessels using a 3-D cylindrical intensity model," *Trans. Img. Proc.*, vol. 16, pp. 1994–2004, Aug 2007.

[67] W. Wong and A. Chung, "Principal curves to extract vessels in 3D angiograms," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, pp. 1–8, 2008.

[68] F. Moselewski, D. Ropers, K. Pohle, U. Hoffmann, M. Ferencik, R. C. Chan, R. C. Cury, S. Abbara, I. kyung Jang, T. J. Brady, W. G. Daniel, and S. Achenbach, "Comparison of measurement of cross-sectional coronary atherosclerotic plaque and vessel areas by 16-slice multidetector computed tomography versus intravascular ultrasound," *The American Journal of Cardiology*, vol. 94, no. 10, pp. 1294–1297, 2004.

[69] X. Li, G. Zhu, T. Yang, X. Xia, K. Fan, Q. Zhang, H. Qiao, S. Li, and D. Li, "An approach for detecting cerebral artery stenosis based on MRA images," in *World Congress on Medical Physics and Biomedical Engineering May 26-31, 2012, Beijing, China*, vol. 39 of *IFMBE Proceedings*, pp. 2091–2094, Springer Berlin Heidelberg, 2013.

[70] H. Kirişli, M. Schaap, C. Metz, A. Dharampal, W. Meijboom, S. Papadopoulou, A. Dedic, K. Nieman, M. de Graaf, M. Meijs, M. Cramer, A. Broersen, S. Cetin, A. Eslami, L. Flórez-Valencia, K. Lor, B. Matuszewski, I. Melki, B. Mohr, I. Öksüz, R. Shahzad, C. Wang, P. Kitslaar, G. Unal, A. Katouzian, M. Orkisz, C. Chen, F. Precioso, L. Najman, S. Masood, D. Ünay, L. van Vliet, R. Moreno, R. Goldenberg, E. Vuçini, G. Krestin, W. Niessen, and T. van Walsum, "Standardized evaluation framework for evaluating coronary artery stenosis detection, stenosis quantification and lumen segmentation algorithms in computed tomography angiography," *Medical Image Analysis*, vol. 17, no. 8, pp. 859–876, 2013.

[71] B. Nilsson and A. Heyden, "A fast algorithm for level set-like active contours," *Pattern Recognition Letters*, vol. 24, no. 9-10, pp. 1331–1337, 2003.

[72] J. Gaztanaga and M. J. Garcia, "Automated analysis of coronary artery disease by computed tomography," *Mount Sinai Journal of*

*Medicine: A Journal of Translational and Personalized Medicine*, vol. 79, no. 2, pp. 295–301, 2012.

[73] C. Biermann, I. Tsiflikas, C. Thomas, B. Kasperek, M. Heuschmid, and C. Claussen, "Evaluation of computer-assisted quantification of carotid artery stenosis," *Journal of Digital Imaging*, vol. 25, no. 2, pp. 250–257, 2012.

[74] B. Kelm, S. Mittal, Y. Zheng, A. Tsymbal, D. Bernhardt, F. Vega-Higuera, S. Zhou, P. Meer, and D. Comaniciu, "Detection, grading and classification of coronary stenoses in computed tomography angiography," in *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2011*, vol. 6893 of *Lecture Notes in Computer Science*, pp. 25–32, Springer Berlin Heidelberg, 2011.

[75] D. Dey, A. Schuhbaeck, Q. Yang, Z. Fan, G. Germano, S. Achenbach, D. Li, and P. Slomka, "Automated coronary artery tracking in contrast-enhanced whole-heart coronary magnetic resonance angiography at 3.0T," *Journal of Cardiovascular Magnetic Resonance*, vol. 15, no. 1, 2013.

[76] S. Mittal, Y. Zheng, B. Georgescu, F. Vega-Higuera, S. K. Zhou, P. Meer, and D. Comaniciu, "Fast automatic detection of calcified coronary lesions in 3D cardiac CT images," in *Machine Learning in Medical Imaging*, pp. 1–9, 2010.

[77] M. Duval, E. Ouzeau, F. Precioso, and B. Matuszewski, "Coronary artery stenoses detection with random forest," in *MICCAI workshop "3D cardiovascular imaging: a MICCAI segmentation challenge"*, 2012.

[78] C. B. Compas, T. Syeda-Mahmood, P. McNeillie, and D. Beymer, "Automatic detection of coronary stenosis in X-ray angiography through spatio-temporal tracking," in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pp. 1299–1302, 2014.

[79] A. Eslami, A. Aboee, Z. Hodaei, M. J. Moghaddam, S. G. Carlier, A. Katouzian, and N. Navab, "Quantification of coronary arterial stenosis by inflating tubes in CT angiographic images," in *MICCAI workshop "3D cardiovascular imaging: a MICCAI segmentation challenge"*, 2012.

[80] J. Zhou, W. Huang, Y. Chi, Y. Duan, L. Zhong, X. Zhao, J. Zhang, W. Xiong, R. S. Tan, and K. K. Toe, "Quantification of coronary artery stenosis by area stenosis from cardiac CT angiography," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 695–698, 2015.

[81] G. Yang, X. Zhao, L. Tang, H. Shu, and C. Toumoulin, "Quantitative measurement of coronary artery stenosis in CCTA images using a 2D parametric intensity model," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1071–1074, 2014.

[82] D. Engelberg and A. Seffah, *A Framework for Rapid Mid-Fidelity Prototyping of Web Sites*, pp. 203–215. Boston, MA: Springer US, 2002.

[83] S. J. Gandy, M. Lambert, J. Belch, I. Cavin, E. Crowe, R. Littleford, J. A. MacFarlane, S. Z. Matthew, P. Martin, R. S. Nicholas, A. Struthers, F. Sullivan, S. A. Waugh, R. D. White, J. R. Weir-McCall, and J. G. Houston, "3T MRI investigation of cardiac left ventricular structure and function in a UK population: The tayside screening for the prevention of cardiac events (TASCFORCE) study," *Journal of Magnetic Resonance Imaging*, vol. 44, pp. 1186–1196, 2016.

[84] M. A. Lambert, J. R. Weir-McCall, M. Salsano, S. J. Gandy, D. Levin, I. Cavin, R. Littleford, J. A. MacFarlane, S. Z. Matthew, R. S. Nicholas, A. D. Struthers, F. Sullivan, S. A. Henderson, R. D. White, J. J. F. Belch, and J. G. Houston, "Prevalence and distribution of atherosclerosis in a low- to intermediate-risk population: Assessment with whole-body MR angiography," *Radiology*, vol. 287, no. 3, pp. 795–804, 2018.

[85] J. R. Weir-McCall, M. Lambert, S. J. Gandy, J. J. F. Belch, I. Cavin, S. A. Henderson, R. Littleford, J. A. Macfarlane, S. Z. Matthew, R. Stephen Nicholas, A. D. Struthers, F. Sullivan, R. D. White, and J. Graeme Houston, "Systemic arteriosclerosis is associated with left ventricular remodeling but not atherosclerosis: a TASCFORCE study," *Journal of Cardiovascular Magnetic Resonance*, vol. 20, no. 1, p. 7, 2018.

[86] C. Wachinger, B. Glocker, J. Zeltner, N. Paragios, N. Komodakis, M. S. Hansen, and N. Navab, "Deformable mosaicing for whole-

body MRI," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008*, pp. 113–121, 2008.

[87] J. Maintz and M. A. Viergever, "A survey of medical image registration," *Medical Image Analysis*, vol. 2, pp. 1–36, 1998.

[88] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: a survey," *IEEE Transactions on Medical Imaging*, vol. 22, pp. 986–1004, 2003.

[89] S. Klein, M. Staring, and J. P. W. Pluim, "Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines," *IEEE Transactions on Image Processing*, vol. 16, pp. 1057–7149, 2007.

[90] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Transactions on Medical Imaging*, vol. 32, pp. 1153–1190, 2013.

[91] M. A. Viergever, J. A. Maintz, S. Klein, K. Murphy, M. Staring, and J. P. Pluim, "A survey of medical image registration – under review," *Medical Image Analysis*, vol. 33, pp. 140–144, 2016.

[92] W. Crum, D. Hill, and D. Hawkes, "Information theoretic similarity measures in non-rigid registration," in *Information Processing in Medical Imaging* (C. Taylor and J. Noble, eds.), vol. 2732 of *Lecture Notes in Computer Science*, pp. 378–387, Springer Berlin Heidelberg, 2003.

[93] W. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, "Multi-modal volume registration by maximization of mutual information," *Medical Image Analysis*, vol. 1, pp. 35–51, 1996.

[94] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Transactions on Medical Imaging*, vol. 16, pp. 187–198, 1997.

[95] J.-P. Thirion, "Image matching as a diffusion process: an analogy with Maxwell's demons," *Medical Image Analysis*, vol. 2, no. 3, pp. 243–260, 1998.

[96] K. Goatman and S. Murphy, "Final report: Registration second look," internal report, Toshiba Medical Visualization Systems, 2012.

[97] M. T. Dehkordi, S. Sadri, and A. Doosthoseini, "A review of coronary vessel segmentation algorithms," *Journal of medical signals and sensors*, vol. 1, no. 1, p. 49, 2011.

[98] C. A. Lupaşcu, D. Tegolo, and E. Trucco, "Accurate estimation of retinal vessel width using bagged decision trees and an extended multiresolution hermite model," *Medical Image Analysis*, vol. 17, no. 8, pp. 1164–1180, 2013.

[99] M. Boccardi, R. Ganzola, M. Bocchetta, M. Pievani, A. Redolfi, G. Bartzokis, R. Camicioli, J. Csernansky, M. de Leon, L. deToledo Morrell, R. Killiany, S. Lehéricy, J. Pantel, J. Pruessner, H. Soininen, C. Watson, S. Duchesne, C. Jack Jr, and G. Frisoni, "Survey of protocols for the manual segmentation of the hippocampus: Preparatory steps towards a joint EADC-ADNI harmonized protocol," *Journal of Alzheimer's Disease*, vol. 26, no. s3, pp. 61–75, 2011.

[100] L. Joskowicz, D. Cohen, N. Caplan, and J. Sosna, "Automatic segmentation variability estimation with segmentation priors," *Medical Image Analysis*, vol. 50, pp. 54–64, 2018.

[101] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, J. Buatti, S. Aylward, J. Miller, S. Pieper, and R. Kikinis, "3D slicer as an image computing platform for the quantitative imaging network," *Magnetic Resonance Imaging*, vol. 30, pp. 1323–41, Nov 2012.

[102] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[103] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, pp. 266–277, Feb 2001.

[104] Y. Zhang, B. J. Matuszewski, L. K. Shark, and C. J. Moore, "Medical image segmentation using new hybrid level-set method," in *2008 Fifth International Conference BioMedical Visualization: Information Visualization in Medical and Biomedical Informatics*, pp. 71–76, July 2008.

[105] A. de Brebisson and G. Montana, "Deep neural networks for anatomical brain segmentation," in *2015 IEEE Conference on*

*Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 20–28, June 2015.

[106] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. N. L. Benders, and I. Išgum, "Automatic segmentation of MR brain images with a convolutional neural network," *IEEE Transactions on Medical Imaging*, vol. 35, pp. 1252–1261, May 2016.

[107] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," *IEEE Transactions on Medical Imaging*, vol. 35, pp. 1240–1251, May 2016.

[108] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[109] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, no. 1, p. 29, 2015.

[110] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, pp. 62–66, Jan 1979.

[111] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[112] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, June 2015.

[113] A. McNeil, G. Degano, I. Poole, G. Houston, and E. Trucco, "Comparison of automatic vessel segmentation techniques for whole body magnetic resonance angiography with limited ground truth data," in *Medical Image Understanding and Analysis*, pp. 144–155, 2017.

[114] F. Milletari, N. Navab, and S. A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, Oct 2016.

[115] G. R. Pinheiro, R. Voltoline, M. Bento, and L. Rittner, "V-net and u-net for ischemic stroke lesion segmentation in a small dataset of perfusion data," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pp. 301–309, 2019.

[116] T. A. Lampert, A. Stumpf, and P. Gancarski, "An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2557–2572, 2016.

[117] P. Jannin, C. Grova, and C. R. Maurer, "Model for defining and reporting reference-based validation protocols in medical image processing," *International Journal of Computer Assisted Radiology and Surgery*, vol. 1, no. 2, pp. 63–73, 2006.

[118] L. Maier-Hein, A. Groch, A. Bartoli, S. Bodenstedt, G. Boissonnat, P. L. Chang, N. T. Clancy, D. S. Elson, S. Haase, E. Heim, J. Hornegger, P. Jannin, H. Kenngott, T. Kilgus, B. Muller-Stich, D. Oladokun, S. Rˆhl, T. R. dos Santos, H. P. Schlemmer, A. Seitel, S. Speidel, M. Wagner, and D. Stoyanov, "Comparative validation of single-shot optical techniques for laparoscopic 3-D surface reconstruction," *IEEE Transactions on Medical Imaging*, vol. 33, no. 10, pp. 1913–1930, 2014.

[119] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy, "Supervised learning from multiple experts: Whom to trust when everyone lies a bit," in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 889–896, 2009.

[120] N. Takaya, J. Cai, M. S. Ferguson MT, V. L. Yarnykh, B. Chu, T. Saam, N. L. Polissar, J. Sherwood, R. C. Cury, R. J. Anders, K. O. Broschat, D. Hinton, K. L. Furie, T. S. Hatsukami, and C. Yuan, "Intra- and interreader reproducibility of magnetic resonance imaging for quantifying the lipid-rich necrotic core is improved with gadolinium contrast enhancement," *Journal of Magnetic Resonance Imaging*, vol. 24, no. 1, pp. 203–210, 2006.

[121] S. J. Gandy, S. A. Waugh, R. S. Nicholas, H. J. Simpson, W. Milne, and J. G. Houston, "Comparison of the reproducibility of quantitative cardiac left ventricular assessments in healthy volunteers using different MRI scanners: A multicenter simulation," *Journal of Magnetic Resonance Imaging*, vol. 28, no. 2, pp. 359–365, 2008.

[122] A. Singh, C. D. Steadman, J. N. Khan, M. A. Horsfield, S. Bekele, S. A. Nazir, P. Kanagala, N. G. Masca, P. Clarysse, and G. P. McCann, "Intertechnique agreement and interstudy reproducibility of strain and diastolic strain rate at 1.5 and 3 tesla: A comparison of feature-tracking and tagging in patients with aortic stenosis," *Journal of Magnetic Resonance Imaging*, vol. 41, no. 4, pp. 1129–1137, 2015.

[123] L. D. Fisher, M. P. Judkins, J. Lesperance, A. Cameron, P. Swaye, T. Ryan, C. Maynard, M. Bourassa, J. W. Kennedy, A. Gosselin, H. Kemp, D. Faxon, L. Wexler, and K. B. Davis, "Reproducibility of coronary arteriographic reading in the coronary artery surgery study (CASS)," *Catheterization and Cardiovascular Diagnosis*, vol. 8, no. 6, pp. 565–575, 1982.

[124] T. Saam, T. S. Hatsukami, V. L. Yarnykh, C. E. Hayes, H. Underhill, B. Chu, N. Takaya, J. Cai, W. S. Kerwin, D. Xu, N. L. Polissar, B. Neradilek, W. K. Hamar, J. Maki, D. W. Shaw, R. J. Buck, B. Wyman, and C. Yuan, "Reader and platform reproducibility for quantitative assessment of carotid atherosclerotic plaque using 1.5T Siemens, Philips, and General Electric scanners," *Journal of Magnetic Resonance Imaging*, vol. 26, no. 2, pp. 344–352, 2007.

[125] R. N. Planken, J. H. Tordoir, R. Dammers, M. W. de Haan, T. K. Oei, F. M. van der Sande, J. M. van Engelshoven, and T. Leiner, "Stenosis detection in forearm hemodialysis arteriovenous fistulae by multiphase contrast-enhanced magnetic resonance angiography: Preliminary experience," *Journal of Magnetic Resonance Imaging*, vol. 17, no. 1, pp. 54–64, 2003.

[126] N. N. Naguib, K. Bohrt, N.-E. A. Nour-Eldin, B. Schulz, A. M. Tawfik, P. Siebenhandel, B. Bodelle, K. Eichler, A. Moritz, T. J. Vogl, and S. Zangos, "Whole-body MR angiography: First experiences with the new TimCT technology with single contrast injection," *Journal of Magnetic Resonance Imaging*, vol. 39, no. 2, pp. 434–439, 2014.

[127] J. Weir-McCall, R. White, P. Ramkumar, S. Gandy, F. Khan, J. Belch1, S. Allan D., and J. Houston, "Follow-up of atheroma burden with sequential whole body contrast enhanced MR angiography: a feasibility study," *Int J Cardiovascular Imaging*, vol. 32, pp. 825–832, 2016.

[128] Q. Chen, Y. Shi, Y. Wang, and X. Li, "Patterns of disease distribution of lower extremity peripheral arterial disease," *Angiology*, vol. 66, no. 3, pp. 211–218, 2015.

[129] K. Krippendorff, "Reliability in content analysis," *Human Communication Research*, vol. 30, no. 3, pp. 411–433, 2004.

[130] L. McCormick, J. Weir-McCall, S. Gandy, R. White, A. McNeil, E. Trucco, and J. G. Houston, "Mapping of arterial location for the design of automated identification and analysis algorithms in whole body MRA," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 7925–7928, Aug 2015.

[131] T. Sandgren, B. Sonesson, A. R. Ahlgren, and T. Lanne, "The diameter of the common femoral artery in healthy human: Influence of sex, age, and body size," *Journal of Vascular Surgery*, vol. 29, no. 3, pp. 503–510, 1999.

[132] T. Sandgren, B. Sonesson, A. R. Ahlgren, and T. Lanne, "Factors predicting the diameter of the popliteal artery in healthy humans," *Journal of Vascular Surgery*, vol. 28, no. 2, p. 284(6), 1998.

[133] M. Sabatier, L. Stoner, M. Reifenberger, and K. Mccully, "Doppler ultrasound assessment of posterior tibial artery size in humans," *Journal of clinical ultrasound*, vol. 34, pp. 223–30, 2006.

[134] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M. A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, Ç. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H. C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. V. Leemput, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.

[135] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[136] L. Gaudette and N. Japkowicz, "Evaluation methods for ordinal classification," in *Advances in Artificial Intelligence*, pp. 207–210, 2009.

[137] A. O'Neil, E. Beveridge, G. Houston, L. McCormick, and I. Poole, "Arterial tree tracking from anatomical landmarks in magnetic resonance angiography scans," in *Medical Imaging 2014: Image Processing*, pp. 708–715, 2014.

[138] R. Shahzad, O. Dzyubachyk, M. Staring, J. Kullberg, L. Johansson, H. Ahlström, B. P. Lelieveldt, and R. J. van der Geest, "Automated extraction and labelling of the arterial tree from whole-body MRA data," *Medical Image Analysis*, vol. 24, no. 1, pp. 28–40, 2015.

[139] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint, arXiv:1409.1556*, 2014.

[140] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[141] Y. Yu, H. Lin, J. Meng, X. Wei, H. Guo, and Z. Zhao, "Deep transfer learning for modality classification of medical images," *Information*, vol. 8, no. 3, p. 91, 2017.

[142] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Artificial Neural Networks and Machine Learning – ICANN 2018*, pp. 270–279, 2018.

[143] A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent, and R. Cipolla, "Understanding real world indoor scenes with synthetic data," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[144] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[145] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[146] E. Richardson, M. Sela, and R. Kimmel, "Understanding real world indoor scenes with synthetic data," in *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 460–469, 2016.

[147] H. Chen, M. Valerio Giuffrida, P. Doerner, and S. A. Tsaftaris, "Adversarial large-scale root gap inpainting," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[148] A. Q. O'Neil, A. Kascenas, J. Henry, D. Wyeth, M. Shepherd, E. Beveridge, L. Clunie, C. Sansom, E. Šeduikytė, K. Muir, and I. Poole, "Attaining human-level performance with atlas location autocontext for anatomical landmark detection in 3D CT data," in *Computer Vision — ECCV 2018 Workshops*, pp. 470–484, 2019.